Navigating Analytical Crossroads:

Charting a course for open, reliable and transparent modelling in ecology and conservation with new frameworks and practical tools

Elliot Gould

B.A., B.Sc., M.Sc.

ORCID 0000-0002-6585-538X

Doctor of Philosophy

A thesis submitted in total fulfillment for the degree of

Doctor of Philosophy

The University of Melbourne

January 2026, revised June 2026

School of Agriculture, Food and Ecosystem Sciences, Faculty of Science

Contents

List of Figures iii

List of Tables iii

List of Listings iii

Abstract v

Declaration of Authorship vii

Preface ix

Acknowledgements xv

1 General Introduction 1

1.1 Reproducibility & Transparency in Ecological Modelling . . . 3

1.1.1 Causes of Unsatisfactory Model Reporting Practices . . . 5

1.1.2 Consequences of Unsatisfactory Model Reporting Practices . . . 7

1.2 Researcher Degrees of Freedom: A Hidden Driver of Irreproducibility . . . 8

1.3 Problem Statement & Thesis Aims . . . 11

2 "Same Data, Different Analysts:" Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology 15

Preface . . . 15

Abstract . . . 16

2.1 Introduction . . . 17

2.2 Methods . . . 21

2.2.1 Step 1: Select Datasets . . . 21

2.2.2 Step 2: Recruitment and Initial Survey of Analysts . . . 24

2.2.3 Step 3: Primary Data Analyses . . . 25

2.2.4 Step 4: Peer Reviews of Analyses . . . 26

2.2.5 Step 5: Evaluate Variation . . . 28

2.2.6 Step 6: Facilitated Discussion and Collaborative Write-Up of Manuscript . . . 42

2.3 Results . . . 42

2.3.1 Summary Statistics . . . 42

2.3.2 Distribution of Effects . . . 45

2.3.3 Quantifying Heterogeneity . . . 47

2.3.4 Post-hoc Analysis: Exploring outlier characteristics and the effect of outlier removal on heterogeneity . . . 51
2.3.5 Post-hoc Analysis: Exploring the effect of removing analyses with poor peer ratings on heterogeneity . . . 53
2.3.6 Post-hoc Analysis: Exploring the effect of including only analyses conducted by analysis teams with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area . . . 55
2.3.7 Post-hoc Analysis: Exploring the effect of excluding estimates of Z_r in which we had reduced confidence . . . 55
2.3.8 Post-hoc Analysis: Exploring the effect of excluding effects from blue tit models that contained two highly collinear predictors . . . 56
2.3.9 Explaining Variation in Deviation Scores . . . 56
2.3.10 Deviation Scores as Explained by Reviewer Ratings . . . 58
2.3.11 Deviation Scores as Explained by the Distinctiveness of Variables in Each Analysis . . . 59
2.3.12 Deviation Scores as Explained by the Inclusion of Random Effects 61
2.3.13 Multivariate Analysis Effect Size (Z_r) and Out-of-sample Predictions (y_i) . . . 62
2.4 Discussion . . . 62
2.5 Conclusions . . . 70

3 *ManyEcoEvo*: an R package for Analysing Many-analyst Style Data 73
Preface . . . 73
3.1 Summary . . . 73
3.1.1 Statement of Need . . . 75
3.2 State of the Field . . . 76
3.3 Research Impact . . . 78
3.4 Software Design . . . 78
3.5 Software Demonstration . . . 81
3.5.1 Dataset Description . . . 81
3.5.2 Calculating meta-analytic results from raw estimates with the blue tit dataset . . . 83
3.6 Scaling up: Systematic Analysis Across Datasets and Analytical Subsets . 88
3.6.1 Preparing to map analyses over datasets with the nested tibble . . 89
3.6.2 Coordinating the Full Pipeline with `meta_analyse_datasets()` 89
3.6.3 Repeating the Analysis Across Multiple Datasets . . . 90
3.6.4 Generating Analytical Subsets . . . 91
3.7 Conclusion . . . 94

4 A Framework for Questionable Research Practices in Ecological Modelling 95
Preface . . . 95
Abstract . . . 95
4.1 Introduction . . . 96
4.2 Conditions for Questionable Research Practices in Ecological Modelling . 99

4.2.1 Identifying motivations for QRPs . . . 99
4.2.2 Which Modelling Objects (Inputs, Outputs, the Model Itself) Are Affected By QRPs? . . . 103
4.3 Mapping QRPs onto the Modelling Process . . . 108
4.3.1 Methods . . . 109
4.3.2 Results . . . 109
4.4 Formalising the "Garden of Forking Paths" in Model-Based Research . . . 118
4.5 A Typology of QRPs . . . 120
4.6 Discussion . . . 125
4.6.1 *Transparency* Determines When 'Poor Practices' are 'Questionable' 125
4.6.2 *Opportunism* Determines When Data-dependent Analytic Decisions are 'Questionable' . . . 126
4.6.3 Data Constraints Amplify the Risk of QRPs . . . 126
4.6.4 Impacts of QRPs . . . 127
4.6.5 Potential Solutions . . . 128
4.6.6 Future Research . . . 131
4.6.7 Conclusion . . . 132

5 "But I Can't Preregister My Research:" Improving the Reproducibility and Transparency of Ecology and Conservation with Adaptive Preregistration for Model-based Research 135
Preface . . . 135
Abstract . . . 136
5.1 Introduction . . . 137
5.2 Materials & Methods . . . 140
5.2.1 Problem Formulation: Defining vision and objectives for a preregistration template for ecological modelling . . . 141
5.2.2 Designing the General Modelling Workflow . . . 142
5.2.3 Designing the Preregistration Template . . . 142
5.2.4 Developing an Adaptive Preregistration Methodology . . . 142
5.2.5 Evaluating and Refining Adaptive Preregistration with a Case Study 144
5.3 Results . . . 148
5.3.1 The Preregistration Template . . . 148
5.3.2 Findings from the Case Study . . . 148
5.3.3 Adaptive Preregistration in Practice: Benefits and challenges . . . 151
5.4 Discussion . . . 159
5.4.1 When is Adaptive Preregistration useful? . . . 160
5.4.2 Difficulties encountered using Adaptive Preregistration in our case study . . . 161
5.4.3 Future Work . . . 164
5.4.4 Conclusions . . . 167

6 General Discussion 169
6.1 Recommendations for Research Practice . . . 172
6.1.1 Improve Precision and Clarity of Research Question Framing . . . 172

6.1.2 Improve Model Transparency to Reduce the Risk of QRPs, Facilitate Model Reuse and Synthesis . . . 174
6.1.3 Refine Adaptive Preregistration Infrastructure & Templates to Facilitate Preregistration Uptake . . . 176
6.2 Contextualising the Work in this Thesis . . . 178
6.3 Conclusion . . . 180

References **183**

Appendices **229**

A Supporting Information for the "Same Data, Different Analysts" Study **229**
A.1 Summarising Variation Among Analysis Specifications . . . 229
A.1.1 Summary Statistics . . . 229
A.1.2 Effect Size Specification Analysis . . . 232
A.2 Effect Size Analysis . . . 235
A.2.1 Meta-analysis . . . 235
A.3 Explaining Variation in Deviation Scores . . . 245
A.3.1 Box-Cox Transformation of Response Variable for Model Fitting . . . 245
A.3.2 Model Convergence and Singularity Problems . . . 245
A.3.3 Deviation Scores as Explained by Reviewer Ratings . . . 246
A.3.4 Deviation Scores as Explained by the Distinctiveness of Variables in Each Analysis . . . 255
A.3.5 Deviation Scores as Explained by the Inclusion of Random Effects 257
A.3.6 Multivariate Analysis . . . 258
A.3.7 Model Summary Metrics for Out-of-sample Predictions y_i . . . 260
A.3.8 Post-hoc Analysis: checking the use of model weights in all models explaining deviation from the meta-analytic mean . . . 264
A.3.9 Model Weight Investigation Findings . . . 265
A.4 Correlation Matrices of Case Study Data . . . 273
A.5 R Package References and Session Info . . . 273

B Published Article for the "Same Data, Different Analysts" Study **275**

C Supporting Information for the "Questionable Research Practices in Ecological Modelling" Chapter **313**
C.1 QRPs Identified from Literature Review . . . 313
C.2 Synthetic Example Code . . . 317

D Supporting Information for the "Adaptive Preregistration" Study **331**
D.1 Modelling Workflow Identification . . . 331
D.1.1 Task 1: Identify Scientific Workflows for Ecological Modelling . . . 331
D.1.2 Task 2: Describe a General Workflow for Ecological Modelling . . . 332
D.2 User Research Workshop Materials . . . 333
D.2.1 Activity 1 – Identifying Modelling Workflows . . . 333

D.2.2 Activity 2 – Identifying Challenges and Solutions to Preregistration for Modelling in Ecology . . . 334
D.3 Adaptive Preregistration User Guide . . . 335
D.4 Ecological Modelling Preregistration Template . . . 397
D.5 Case Study Preregistration . . . 416
D.6 Case Study Pilot Analysis Summary Report . . . 448
D.6.1 Pilot Analysis of VEFMAP Riparian Vegetation Surveys . . . 448
D.6.2 Vegetation Cover: regime model . . . 450
D.6.3 Vegetation Cover: flow event model (version 1) . . . 454
D.6.4 Vegetation Cover: flow event model (version 2) . . . 455
D.6.5 Species Richness Model . . . 459
D.7 Case Study Preregistration with git and GitHub . . . 464

E Published Article for the "Adaptive Preregistration" Study **467**

List of Figures

2.1 Schematic of research process showing recruited analyst and reviewer contributions in orange and core team contributions in blue. . . . 28

2.2 Forest plots of meta-analytic estimated standardised effect sizes (Z_r, blue triangles) and their 95% confidence intervals for each effect size included in the meta-analysis model. The meta-analytic mean effect size is denoted by a black circle and a dashed vertical line, with error bars also representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0, indicating no relationship between the test variable and the response variable. Note that the *Eucalyptus* plot omits one extreme outlier with the value of -4.47 (Figure A.2) in order to standardize the x-axes on these two panels. . . . 46

2.3 Forest plot of meta-analytic estimated standardised blue tit out-of-sample predictions, y_i. . . . 48

2.4 Forest plot of meta-analytic estimated *Eucalyptus* out-of-sample predictions, y_i, on the response-scale (stems counts). . . . 49

2.5 Violin plot of Box-Cox transformed deviation from meta-analytic mean Z_r as a function of categorical peer rating. . . . 59

2.6 Fitted model of the Box-Cox transformed deviation score (deviation in effect size from meta-analytic mean) as a function of the mean Sorensen's index showing distinctiveness of the set of predictor variables. Grey ribbons on predicted values are 95%CI's. 60
2.7 Violin plot of Box-Cox transformed deviation from meta-analytic mean as a function of random-effects inclusion in *Eucalyptus* analyses. 61

3.1 Forest plots of meta-analytic estimated standardised effect sizes for a subset of the blue tit dataset. 87
3.2 Scatterplot of Box-Cox transformed deviation scores as a function of the the mean Sorensen's index. 88

4.1 Schematic of the model development process underpinning mathematical formalism. 105
4.2 Synthesis of questionable research practices (QRPs) in ecological modelling. 111
4.3 Illustration of model fishing and scenario hacking (HARKing) with a synthetic example. Extended caption is located in Box 1. 117

5.1 Schematic of process for translating preregistration to ecological modelling research. 146
5.2 Process for implementing Adaptive Preregistration. 147
5.3 Evolution of model development across preregistration iterations in the case study. 152
5.4 Application of registered flexibility within the case study preregistration, for selecting model structure (a) and selecting an alternative functional form over the default zero-inflated Poisson model (b). 158

6.1 Cuevas Shaw's (2022) strategy for change required to foster a culture of open, transparent, and reproducible science, adapted to include a 'Problem Framing' foundation that precedes the cultural change process and incorporates contributions from this thesis. 179

A.1 Specification curve plot for blue tit analyses. 233
A.2 Specification curve plot for *Eucalyptus* analyses. 234
A.3 Orchard plot of meta-analytic model fitted to all *Eucalyptus* analyses. . . 235
A.4 Orchard plot of meta-analytic model fitted to all blue tit analyses 236

A.6 237

A.6 Forest plots of meta-analytic estimated standardised effect sizes for data subsets. 238

A.7 Estimated meta-analytic effect size (Z_r, blue triangles), standard error, and 95% confidence intervals, from analyses of the primary data set with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area. Meta-analytic mean effect sizes are represented with black circles. 241

A.8 Forest plot of meta-analytic estimated effect-sizes standard error and 95% confidence intervals of blue tit analyses with highly collinear analyses removed. 242

A.9 Forest plot of meta-analytic estimated out-of-sample predictions on the response-scale (stem counts) for *Eucalyptus* analyses. 244

A.10 Box-Cox transformed absolute deviation scores plotted against (untransformed) absolute deviation scores. 245

A.11 Violin plot of Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of categorical peer-review ratings ratings. . 251

A.12 Scatterplots explaining Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of continuous peer-review ratings. 252

A.13 Scatter plots examining Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of Sorensen's similarity index.256

A.14 Violin plot of Box-Cox transformed deviation from meta-analytic mean as a function of presence or absence of random effects in the analyst's model.258

A.15 Effect plots for each non-singular model that converged with estimable fixed effect group means for the *Eucalyptus* dataset. 267

A.16 Effect plots for each non-singular model that converged with estimable fixed effect group means for the blue tit dataset. 268

A.17 Model performance comparison plots for a subset of models that passed convergence and singularity checks and had estimable marginal effects. Models are compared based on their performance metrics, including $R^2_{\text{Conditional}}$, R^2_{Marginal}, Intra-Class Correlation, Root Mean Squared Error (RMSE), and the weighted AIC, corrected AIC and BIC. Values of performance are normalised across models for each metric, against the top-most performing model across all metrics. Greater distance from the centre on each metric axis indicates dominance in performance. For both blue tit and *Eucalyptus* models, all included a random effect for Reviewer ID, and weights consisted of the inverse Box-Cox variance for weights (blue line), inverse folded variance (yellow line), or none (red line). 270

D.1 Effects plot for vegetation cover flow regime model. 452

D.2 Model predicted plant cover for the three way interaction between `Days above baseflow`, `functional group` and `origin`. 453

D.3 Model predicted plant cover for the three way interaction between `Days above springfresh`, `functional group` and `origin`. 453

D.4 Vegetation cover flow event model (version 1) effects plot. 455

D.5 Model predicted plant cover for the two way interaction between `period` and `zone`. 456

D.6 Vegetation cover, flow event model (version 2) effects plot. 457

D.7 Model predicted plant cover for the two way interaction between `period` and `functional group`. 458

D.8 Model predicted species richness for the two-way interaction between `period` and `zone`. 461

D.9 Model predicted species richness for the two-way interaction between `zone` and `functional group`. 462

D.10 Model predicted species richness for the two-way interaction between `period` and `functional group`. 463

D.11 Screenshot of GitHub issue thread illustrating where the preregistered analysis tasks and rationale were transparently recorded. 464

D.12 Screenshot of comment on GitHub issue illustrating data-dependent decision-making and where deviation from preregistration occurred during the case study. 465

List of Tables

2.1 Tallies of analysts' qualitative answers to the research questions addressed by their analyses. 45

2.2 Heterogeneity in the estimated effects Z_r for meta-analyses of: the full dataset, as well as from post hoc analyses wherein analyses with outliers are removed, analyses with effects from analysis teams with at least one "unpublishable" rating are excluded, analyses receiving at least one "major revisions" rating or worse excluded, analyses from teams with at least one analyst self-rated as "highly proficient" or "expert" in statistical analysis are included, and (blue tit only) analyses that did not included the pair of highly collinear predictors together. τ^2_{Team} is the absolute heterogeneity for the random effect Team. τ^2_{effectID} is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{\text{Team, effectID}}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences. 50

2.3 Heterogeneity among the out-of-sample predictions y_i for both blue tit and *Eucalyptus* datasets. τ^2_{Team} is the absolute heterogeneity for the random effect Team. τ^2_{effectID} is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{\text{Team,effectID}}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences. 52

2.4 Estimated mean value of the standardised correlation coefficient, Z_r, along with its standard error and 95% confidence intervals. We recomputed the meta-analysis for different post-hoc subsets of the data: All eligible effects, removal of effects from blue tit analyses that contained a pair of highly collinear predictor variables, removal of effects from analysis teams that received at least one peer rating of “deeply flawed and unpublishable”, removal of any effects from analysis teams that received at least one peer rating of either “deeply flawed and unpublishable” or “publishable with major revisions”, inclusion of only effects from analysis teams that included at least one member who rated themselves as “highly proficient” or “expert” at conducting statistical analyses in their research area. 54

2.5 Summary metrics for registered models seeking to explain deviation (Box-Cox transformed absolute deviation scores) from the mean Z_r as a function of Sorensen's index, categorical peer ratings, and continuous peer ratings for blue tit and *Eucalyptus* analyses, and as a function of the presence or absence of random effects (in the analyst's models) for *Eucalyptus* analyses. We report coefficient of determination, R^2, for our models including only fixed effects as predictors of deviation, and we report $R^2_{\text{Conditional}}$, R^2_{Marginal} and the intra-class correlation (ICC) from our models that included both fixed and random effects. For all our models, we calculated the residual standard deviation σ and root mean squared error (RMSE). 56

2.6 Parameter estimates from models of Box-Cox transformed deviation scores from the mean Z_r as a function of continuous and categorical peer ratings, Sorensen scores, and the inclusion of random effects. Standard Errors (SE), 95% confidence intervals (95% CI) are reported for all estimates, while t values, degrees of freedom and *p*-values are presented for fixed-effects. Note that positive parameter estimates mean that as the predictor variable increases, so does the absolute value of the deviation from the meta-analytic mean. 57

4.1 Examples and self-reported frequency of questionable research practices (QRPs) in hypothesis-testing research in ecology and evolutionary biology. QRPs are categorised as "cherry-picking," "*p*-hacking," and "methodologically flawed," indicated by the cherry, saw, and cross icons respectively. Data compiled from Fraser et al. (2018) and abbreviations defined by Makel et al. (2021). Ecology: $n = 494$; Evolution: $n = 313$. 98

4.2 Distinguishing between questionable and defensible motivations for decision making during modelling, and the target of the decisions. I designate questionable practices in grey shaded cells and assign practices to classes of QRPs (described in Table 4.3). See Box 2 for notation and expanded definitions. 122

4.3 Formal descriptions of QRP classes, their definitions and some practical examples. 124

5.1 Key stages of the modelling and Adaptive Preregistration process implemented in the environmental flows modelling case study. Version numbers in parentheses are GitHub tags, marking snapshots of the repository for each Preregistration Version (github.com/egouldo/VEFMAP_VEG_Stage6/releases, osf.io/sq5j4/files/jdbr7). 149

A.1 Summary of the number of analysis teams, total analyses, models with normal error distributions, mixed effects models, and models developed with Bayesian statistical methods for effect size analyses only (Z_r) and out-of-sample prediction only (y_i). 230

A.2 Mean, standard deviation and range of number of fixed and random variables, interaction terms used in models and analysis sample size (N). Repeated for effect-size analyses only (Z_r) and out-of-sample predictions only (y_i). 231

A.3 Mean, SE, minimum and maximum number of analyses in which each variable was used, for effect size analyses only (Z_r), out-of-sample predictions only (y_i), using the full dataset. 231

A.4 Estimated meta-analytic mean, standard error, and 95% confidence intervals, from analyses of the primary data set, the more conservative version of the dataset which excluded effects based on less reliable estimates of df, and both of these datasets with outliers removed. 239

A.5 Analyst-constructed variables and their inclusion in meta-analyses of out-of-sample predictions, y_i. 243

A.6 Singularity and convergence checking outcomes for models of deviation in effect-sizes Z_r explained by peer-review ratings for different random effect structures. Problematic checking outcomes are highlighted in red. 248

A.7 Singularity and convergence checking outcomes for models of deviation in out-of-sample predictions y_i explained by peer-review ratings for different random effect structures. Problematic checking outcomes are highlighted in red. 248

A.8 Singularity and convergence checking for models of deviation in out-of-sample-predictions y_i explained by peer-ratings. 250

A.9 Parameter estimates for univariate models of Box-Cox transformed deviation from the mean y_i estimate as a function of categorical peer-review rating, continuous peer-review rating, and Sorensen's index for blue tit and *Eucalyptus* analyses, and also for the inclusion of random effects for *Eucalyptus* analyses. 253

A.10 Singularity and convergence checks for models of deviation explained by Sorensen's similarity index and inclusion of random effects for out-of-sample predictions, y_i. Models of Deviation explained by inclusion of random effects are not presented for blue tit analyses because the number of models not using random effects was less than our preregistered threshold. 257

A.11 Parameter estimates from models explaining Box-Cox transformed deviation scores from the mean Z_r as a function of continuous and categorical peer-review ratings in multivariate analyses. Standard Errors (SE), 95% confidence intervals (95% CI) are reported for all estimates, while t values, degrees of freedom (df) and p-values are presented for fixed-effects. . . 259

A.12 Model summary metrics for multivariate meta-analysis models. σ is the residual standard deviation, ICC is the intra-class correlation coefficient, and R^2_M and R^2_C are the marginal and conditional R^2, respectively. 259

A.13 Singularity and convergence checks for all combinations of random effects specifications trialled for across subsets of out-of-sample predictions y_i from multivariate models. 260

A.14 Parameter estimates for converging, non-singular multivariate models fitted to blue tit out-of-sample-prediction estimates y_i. 261

A.15 Model summary statistics for non-singular, converging multivariate models fit to out-of-sample estimates y_i. 262

A.16 Model summary metrics for models of Box-Cox transformed deviation from the mean y_i estimate as a function of categorical peer-review rating, continuous peer-review rating, and Sorensen's index for blue tit and *Eucalyptus* analyses, and also for the inclusion of random effects for *Eucalyptus* analyses. Coefficient of determination, R^2, is reported for models of deviation as a function of Sorensen diversity scores and presence of random effects, while $R^2_{\text{Conditional}}$, R^2_{Marginal} and the intra-class correlation (ICC) are reported for models of deviation as explained by peer-review ratings. For all models the residual standard deviation σ, root mean squared error (RMSE) were calculated. The number of observations ($N_{\text{Obs.}}$) is displayed for reference. 263

A.17 Singularity and convergence checks for all combinations of model weights and random-effects structure in models of the effect of categorical peer rating on deviation from the analytic mean. For some models, mean estimates of parameter levels for peer-review rating were not able to be estimated. 266

A.18 Marginal means estimated across weight and random effects specifications, for all estimable models, for both *Eucalyptus* and blue tit datasets. 271

A.19 Model performance metric values (non-normalised) for final subset of models considered in weights analysis. All models in final subset included random-effect of Reviewer ID. Metrics included $R^2_{\text{Conditional}}$, R^2_{Marginal}, Intra-Class Correlation, Root Mean Squared Error (RMSE), and the weighted AIC, corrected AIC and BIC. 272

A.20 R packages used to generate this manuscript. Please see the *ManyEcoEvo* package for a full list of packages used in the analysis pipeline. 273

A.20 R packages used to generate this manuscript. Please see the *ManyEcoEvo* package for a full list of packages used in the analysis pipeline. 274

A.21 Session information for results generated with the 'Same Data, Different Analysts' quarto manuscript (Gould et al., 2024). 274

C.1 Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1. 314
C.1 Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1. 315
C.1 Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1. 316

D.1 Vegetation cover flow regime model summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1. 450
D.1 Vegetation cover flow regime model summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1. 451
D.2 Vegetation cover flow event model (version 1) summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1. 454
D.3 Vegetation cover, flow event model (version 2) summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1. 457
D.4 Species richness model summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1. 459

List of Listings

A.1 Function to calculate the variance of the Box-Cox transformed deviation scores. 265

C.1 Code used to generate synthetic worked example in Figure 4.3. 317

D.1 The preferred full model structure (in high-level, R notation). 448
D.2 Vegetation cover, regime model structure. 452
D.3 Vegetation cover, flow event model (version 1). 454
D.4 Vegetation cover, flow event model (version 2). 456
D.5 Species richness model structure. 459

Abstract

The reproducibility crisis documented across psychology, medicine, and other scientific disciplines has prompted critical examination of research practices in ecology. While initial meta-research suggests ecology faces similar risks, the specific challenges confronting ecological modelling remain underexplored. The role of "researcher degrees of freedom," or the numerous subjective decisions researchers make throughout the scientific process, including modelling, is only recently being acknowledged as central to reproducibility concerns across disciplines. When these analytic choices remain undisclosed, inadequately documented, or exploited, they can undermine the reliability and reproducibility of model-based insights. Despite growing calls for transparency in ecological research, the extent to which reproducibility challenges manifest in ecological modelling has received limited systematic investigation. Given the critical role of ecological models in underpinning conservation decision-making, it is essential and timely that research effort is directed at exploring this issue.

The overarching goal of this thesis is to investigate the reproducibility and transparency of ecological modelling, with a focus on applied research. In four empirical chapters, I examine how researcher degrees of freedom threaten research credibility whilst exploring solutions that preserve methodological flexibility essential for complex modelling. In Chapter 2, I present the first large-scale many-analyst study in ecology and evolutionary biology, which revealed numerous researcher degrees of freedom, with analytic flexibility substantially altering inferences and conclusions. To facilitate reproduction and replication, I developed the ManyEcoEvo R package, creating a fully reproducible analysis pipeline that enables verification of our findings and application to new datasets (Chapter 3).

The considerable analytic variability documented in Chapter 2 is particularly concerning when considered alongside a culture of poor model reporting, which obscures analytical decisions and creates opportunities for questionable research practices (QRPs) to emerge.

In Chapter 4, I address these critical conceptual gaps by extending frameworks for understanding QRPs in the ecological modelling context. I provide the first definition of QRPs applicable to ecological modelling and identify modelling-specific practices for modellers to be wary of.

Recognising that awareness alone cannot prevent QRPs, I developed Adaptive Preregistration (Chapter 5), a structured, principled methodology that constrains researcher degrees of freedom whilst accommodating the data-dependent nature of ecological modelling. Through registered flexibility and interim preregistrations, this approach mitigates QRP risks while preserving analytical adaptability and increasing transparency. A comprehensive preregistration template and real-world case study demonstrate the practical feasibility of this methodology and provide tools for the community to preregister their own model-based research.

This thesis establishes that ecological modelling faces distinct reproducibility challenges requiring discipline-specific solutions. The combination of extensive researcher degrees of freedom and poor transparency creates conditions for a reproducibility crisis that threatens the credibility of model-based research. However, this crisis remains largely unacknowledged within the ecological modelling community. By providing empirical evidence, conceptual frameworks, and practical tools, this work contributes foundational elements necessary for cultural change towards more transparent and reproducible ecological modelling. As environmental challenges intensify, ensuring the reliability and transparency of models underpinning environmental decisions is increasingly critical for effective ecological management and public trust in science.

Declaration of Authorship

I, Elliot Gould, declare that this thesis titled “Navigating Analytical Crossroads: Charting a course for open, reliable and transparent modelling in ecology and conservation with new frameworks and practical tools” and the work presented in it are my own. I confirm that:

- the thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is no longer than the maximum word limit in length, exclusive of tables, maps, bibliographies, and appendices.

Signed:

Date:

Preface

The work presented in this thesis is predominately my own, produced under the supervision of Dr Libby Rumpff, Prof Fiona Fidler, Dr Hannah Fraser, and Dr Bonnie Wintle.

Some parts of the work were performed in collaboration with others, as detailed below. No work in this thesis has been submitted for other qualifications or carried out prior to my enrolment in this degree. No digital editorial assistance or generative AI tools were used in preparing the thesis. No third party editorial assistance was provided in preparation of the thesis.

This thesis is a 'thesis with publications' in which each chapter is written in the format of a journal article. Each empirical chapter is a distinct component with its own introduction and conclusion sections such that there is some overlap with the general introduction and discussion chapters.

This thesis includes one original paper published in a peer reviewed journal, and one accepted publication, pending revisions. Two further chapters are in preparation for submission. Below I outline the publication status of each chapter and summarise the contributions made by myself and my collaborators as CRediT contribution statements.

Chapter Publication Status and Summary of Contributions

Chapter 2

The contents of Chapter 2 correspond to the author-accepted version of the following published article:

> **Gould, E.**, Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., *et al.* Same data, different analysts: variation in

effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biol* **23**, 35 (2025). https://doi.org/10.1186/s12915-024-02101-x

This chapter is a "multi-analyst" study, comprising a project lead team (THP, HF, SN, EG, SG, PV, DH, FF) and 246 analysts (see Appendix B for full author list). EG, HF and THP are joint first-authors.

HF, THP and FF conceptualised the project. PV provided raw data for Eucalyptus analyses and SG and THP provided raw data for blue tit analyses. DGH, HF and THP prepared surveys for collecting participating analysts and reviewer's data. EG, HF, THP, PV, SN and FF planned the analyses of the data provided by our analysts and reviewers. EG, HF, DGH and THP curated the data, EG and HF wrote the software code to implement the analyses and prepare data visualisations. EG ensured that analyses were documented and reproducible. THP and HF administered the project, including coordinating with analysts and reviewers. FF provided funding for the project. THP, HF, and EG wrote the manuscript. Authors listed alphabetically contributed analyses of the primary datasets or reviews of analyses. All authors read and approved the final manuscript.

Chapter 3

Chapter 3 contains material archived on Zenodo but not yet submitted for publication:

Gould, E., Fraser, H. S., Nakagawa, S., & Parker, T. H. (2026). ManyEcoEvo: an R package for analysing many-analyst style data (v2.8.9000). *Zenodo*. https://doi.org/10.5281/zenodo.10046152

EG conceptualised and conducted the formal analysis presented in the manuscript. EG, HF, THP curated the data, EG and THP validated the data. EG wrote the software with code contributions from HF and SN. EG prepared data visualisations with contributions from HF. EG wrote the original draft manuscript and software documentation. EG and HF reviewed and edited the final manuscript.

Chapter 4

Chapter 4 contains preprinted material not yet submitted for publication:

Gould, E., Fraser, H. S., Wintle, B. C., Rumpff, L., & Fidler, F. (2026). A Framework for Questionable Research Practices in Ecological Modelling. Applied Statistics. https://doi.org/10.32942/X2DQ0K

EG, HF, FF conceptualised this study. EG curated the data and conducted the formal analysis (literature review, data simulation and visualisation, development of mathematical formalism). EG and HF planned the data analysis. EG wrote the data simulation and visualisation code. Visualisations were prepared by EG and BW, with support from HF. EG was supervised by HF, FF, LR and BW. EG wrote the original draft manuscript. All authors edited and reviewed the manuscript.

Chapter 5

The contents of Chapter 5 correspond to the author-accepted version of the following published article:

Gould, E., Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H. F., Good, M. K., Duncan, D. H., Hauser, C. E., Wintle, B. C., & Rumpff, L. (2026). 'But I can't preregister my research': Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research. Methods in Ecology and Evolution, 2041-210x.70311. https://doi.org/10.1111/2041-210x.70311

EG, CJ, JDLY, HF, and LR conceptualised the project. EG, LR, and CJ acquired or provided funding. EG and CJ administered the project. EG, CJ, JDY, and HW curated the data. EG, CJ, HF, CEH participated in the collaborative workshop, which was planned by EG, HF, LR, CJ and facilitated by EG, CJ, HF. EG, CJ, JDLY, LR, DHD, MKG contributed to the design and refinement of the preregistration template and Adaptive Preregistration methodology. EG analysed the workshop outputs, and JDLY, HW, CJ conducted the case study modelling, with support from EG. JDLY, HW, CJ and EG wrote the code to analyse the workshop outputs and implement the case study modelling and data analysis. EG, JDLY, and CJ validated the data and ensured the analysis was reproducible. EG, CJ, JDLY and BW prepared data visualisations. EG was supervised by LR, HF, and BW. EG wrote the original draft manuscript with support from BW. All authors reviewed and edited the final manuscript. EG wrote the

Adaptive Preregistration website and guide, which was reviewed and edited by HF, LR, and CEH.

Additional Peer-reviewed Publications & Reports

Additional publications and reports arising from research undertaken alongside my candidature are listed below.

- Ivimey-Cook, E. R., Sánchez-Tójar, A., Berberi, I., Čulina, A., Roche, D. G., Almeida, R. A., Amin, B., Bairos-Novak, K. R., Balti, H., Bertram, M., Bliard, L., Byrne, I., Chan, Y.-C., Cioffi, W. R., Corbel, Q., Elsy, A. D., Florko, K. R. N., **Gould, E.**, Grainger, M., … Moran, N. P. (2025). From Policy to Practice: Progress towards Data- and Code-Sharing in Ecology and Evolution. *Proceedings of the Royal Society (B)*. https://doi.org/10.1098/rspb.2025.1394

- Good, M. K., Rumpff, L., Fraser, H., **Gould, E.**, Jones, C. S., Prober, S. M., Bourne, M., Butt, N., Byrne, M., Duncan, D. H., Gorrod, E., Gosper, C. R., Jordan, R., McIntyre, S., Moore, J. L., Nerenberg, S., Pulsford, S., Richards, A. E., Rogers, D. J., … Vesk, P. A. (2024). A structured approach for building multi-community State and Transition Models to support conservation planning. *Journal of Applied Ecology*, *61*(9), 2294–2307. https://doi.org/10.1111/1365-2664.14718

- Nakagawa, S., Ivimey-Cook, E. R., Grainger, M. J., O'Dea, R. E., Burke, S., Drobniak, S. M., **Gould, E.**, Macartney, E. L., Martinig, A. R., Morrison, K., Paquet, M., Pick, J. L., Pottier, P., Ricolfi, L., Wilkinson, D. P., Willcox, A., Williams, C., Wilson, L. A. B., Windecker, S. M., … Lagisz, M. (2023). Method Reporting with Initials for Transparency (MeRIT) promotes more granularity and accountability for author contributions. *Nature Communications*, *14*(1), 1788. https://doi.org/10.1038/s41467-023-37039-1

- Jones, C. S., Thomas, F. M., Michael, D. R., Fraser, H., **Gould, E.**, Begley, J., Wilson, J., Vesk, P. A., & Rumpff, L. (2023). What state of the world are we in? Targeted monitoring to detect transitions in vegetation restoration projects. *Ecological Applications*, *33*(1), e2728. https://doi.org/10.1002/eap.2728

- Ivimey-Cook, E. R., Pick, J. L., Bairos, K. R., Čulina, A., **Gould, E.**, Grainger, M., Marshall, B. M., Moreau, D., Paquet, M., Royauté, R., Sánchez, A., Silva, I., & Windecker, S. M. (2023). Implementing code review in the scientific workflow:

Insights from ecology and evolutionary biology. *Journal of Evolutionary Biology*. https://doi.org/10.1111/jeb.14230

- Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., **Gould, E.**, Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P., Pearson, R., Singleton Thorn, F., Ashton, R., Willcox, A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., … Fidler, F. (2023). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLOS ONE*, *18*(1), e0274429. https://doi.org/10.1371/journal.pone.0274429

- Wintle, B. C., Smith, E. T., Bush, M., Mody, F., Wilkinson, D. P., Hanea, A. M., Marcoci, A., Fraser, H., Hemming, V., Thorn, F. S., McBride, M. F., **Gould, E.**, Head, A., Hamilton, D. G., Kambouris, S., Rumpff, L., Hoekstra, R., Burgman, M. A., & Fidler, F. (2023). Predicting and reasoning about replicability using structured groups. *Royal Society Open Science*. https://doi.org/10.1098/rsos.221553

- O'Dea, R. E., Parker, T. H., Chee, Y. E., Čulina, A., Drobniak, S. M., Duncan, D. H., Fidler, F., **Gould, E.**, Ihle, M., Kelly, C. D., Lagisz, M., Roche, D. G., Sánchez-Tójar, A., Wilkinson, D. P., Wintle, B. C., & Nakagawa, S. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*(1). https://doi.org/10.1186/s12915-021-01006-3

- Hanea, A. M., Wilkinson, D. P., McBride, M., Lyon, A., Van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., **Gould, E.**, Smith, E. T., Mody, F., Bush, M., Fidler, F., Fraser, H., & Wintle, B. C. (2021). Mathematically aggregating experts' predictions of possible futures. *PLOS ONE*, *16*(9), e0256919. https://doi.org/10.1371/journal.pone.0256919

- Jones, C., Vivian, L., Wootton, H., **Gould, E.**, Yen, J., Mole, B., Thomas, F., Backstrom, A., and Just, K. (2025). Vegetation responses to managed river flow events and regimes. *Arthur Rylah Institute for Environmental Research [Unpublished Report for the Department of Energy, Environment and Climate Action]*, Victoria.

- Good, M., Fraser, H., **Gould, E.**, Vesk, P., Rumpff, L., (2021). A practical guide for conservation planning using the General Ecosystem Model for Southern Australian Woodlands. *NESP Threatened Sprecies Recovery Hub Project 7.2, Brisbane*. https://www.nespthreatenedspecies.edu.au/media/yvlbs1tk/7-2-a-

practical-guide-for-conservation-planning-using-the-general-ecosystem-model-for-southern-australian-woodlands_v3.pdf

Copyright notice

© Elliot Gould (2026).

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Appendix B contains the published article of Chapter 2, reproduced in entirety with permission under license CC BY-NC-ND 4.0. Appendix E contains the published article of Chapter 5, reproduced in entirety with permission under license CC BY-NC 4.0.

The remainder of this thesis is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Acknowledgements

First and foremost, I extend my deepest thanks to all my supervisors, who gave me tremendous freedom for creativity and exploration during this PhD. It has been an absolute delight to work with you all.

Libby, I am so grateful you were willing to supervise me a second time around! Your mentorship has helped me grow in confidence in my ideas, skills, and abilities bit by bit. Thank you for always 'pumping up my tyres,' especially over the final stretch. I am deeply appreciative of always having your support. Your perspective has guided me through countless challenges and decisions over the years. Fiona, your expertise in helping me bridge diverse literatures and wrangle a mess of sometimes incongruent concepts and terminologies from different schools of thought has been invaluable. I am also grateful for the many opportunities to travel and present at conferences, and for connecting me to colleagues across the field. These experiences were truly influential in shaping the trajectory of my work and have helped me foster a global collegial network. Hannah, your grounded approach to research and problem-solving always helped me zoom out and see the bigger picture, cutting through the noise with practical and thoughtful solutions. I am also grateful for our work together on the many-analysts project — it was a pleasure to have you as a co-author throughout. Bonnie, you were a late but very welcome addition to the team. I had so much fun with you crafting the writing and arguments in the final chapters; your editorial guidance transformed what felt like chaos into neat and concise pieces of work. To the remainder of my supervisory committee, Peter Vesk and Cindy Hauser, thank you for your rigorous critiques and support of my ideas. Pete, I particularly appreciate your door being open for check-ins to work out a way forward through various sticking points.

To my collaborators: Chris Jones and Jian Yen, I am grateful for your patience, for letting me gatecrash the VEFMAP study and for trusting me to give preregistration a go. To Tim Parker, meeting with you and Hannah every Friday morning was the highlight of my time

during Melbourne's never-ending lockdowns and something I looked forward to immensely. It was a pleasure to collaborate with you.

To my colleagues: Thank you Rose O'Dea for your mentorship, company, and advice, shared over weekly tennis games. David Wilkinson, thank you for being an invaluable sounding board and wonderful collaborator on the repliCATS project. Steven Kambouris, your work on reproducibility has been surprisingly helpful in tying the threads of the thesis together. I'm very grateful for your encouraging text messages at the finish line, and for showing me that reaching the end is possible!

Thank you to my QAECO x ERA colleagues and friends: Saoirse Kelleher, for your thoughtful discussions on all things SORTEE and EcoEvo, Eleanor Stern for your encouragement and proofreading, Ainsley Macdonald for your company and motivation to shut-up-and-write, and to Shona for your warm friendship, care and inclusivity that made me feel welcomed and connected.

Finally, to my family and friends. I wish to express my deepest gratitude to Daniella Segal, who began this journey as my partner and has since become a lifelong friend. Thank you for taking on all the responsibilities at home, without a second thought, so I could continue working on the thesis whilst recovering with long-COVID, and for taking me in as a housemate again over these last few months to see this thesis to completion!

To Charle, the asymptote is finally destroyed! But only because of your love, care and patience. I'm grateful for your proofreading, your tedious BibTeX fixes and care packages over the last year.

To the wonderful Melburn Durt crew, especially Aroha, RJ, Kia, Jackie, Meg, and Em: Thank you for accompanying me to the hospital and for looking after me in recovery after all my bike crashes and health challenges over the years. Your meal and snack drops made me feel cared for and allowed me to focus on rest and recovery.

To my brother Declan, thank you for your time helping me with those fiddly QA checks and your kind words of support. To my sister Rachelle: For being my greatest ally, I'm so grateful that you always have my back.

Last, but not least, to my mum Liane and dad Peter: Thank you for always valuing and encouraging my education. Dad, I appreciate the open-mindedness, critical thought, and

curiosity that you fostered in me from a young age, and for becoming my favourite 'flatmate-from-hell' over these last few months. Mum, you showed me the meaning of perseverance, resilience, and strength in the face of life's challenges — I could not have made it through without your tenacity lighting the way.

Funding Support

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220 and additional funding sources:

- "AIMOS top-up scholarship," 2022, *Association for Interdisciplinary Metaresearch and Open Science (AIMOS)*.
- "Melbourne Centre of Data Science Doctoral Academy Fellow," 2021, *Melbourne Centre for Data Science, University of Melbourne*.
- "Testing and Developing Preregistration Templates for Ecology and Conservation Using a Case Study of Environmental Flows Management in Victoria, Australia," 2020, *Arthur Rylah Institute, Department of Energy, Environment and Climate Action*.

I have been supported in attending conferences and workshops with the following travel scholarships and grants:

- ESA Travel Grant, 2025, *Ecological Society of Australia Student (ESA) Annual Conference, Adelaide*.
- Science Abroad Travelling Scholarships, 2023, *University of Melbourne, Faculty of Science, School of Ecosystem and Forest Sciences*.
- Metascience Travel Award, 2023, *Metascience Conference 2023, Washington D.C.*
- Travel Grant, 2023, *Centre de la Science de la Biodiversité du Québec at the Université du Québec à Montréal*.

1 General Introduction

Over the past decade, multiple scientific disciplines have confronted evidence of a "reproducibility crisis," where published findings prove difficult or impossible to replicate (Baker, 2016; Camerer et al., 2018; Open Science Collaboration, 2015). This crisis has been documented and provoked heated debate across psychology, medicine, economics and neuroscience (Munafò et al., 2017; Nosek et al., 2021). Although ecology has not expended the same degree of effort in other disciplines in characterising or quantifying the extent of reproducibility issues that it faces (Nakagawa & Parker, 2015), initial assessment of the conditions found to foster reproducibility problems provide evidence that ecology as a discipline is also at risk of a "reproducibility crisis" (Fidler et al., 2017). Self-report surveys suggest that rates of questionable research practices (QRPs) are similar to those in psychology (Fraser et al., 2018) and meta-analytic evidence indicates widespread publication bias and effect-size exaggeration (Deressa et al., 2023; Kimmel et al., 2023; Parker & Yang, 2023; Yang et al., 2022, 2023).

Self-examination in ecology has followed metaresearch efforts in other disciplines, with calls to embrace open science in ecology and conservation (Burgman et al., 2023; Lewis et al., 2023; Powers & Hampton, 2018; Roche et al., 2021; Stewart Lowndes et al., 2017). However, to date most efforts have been restricted to practices that foster *computational reproducibility*, including code and data sharing practices (Čulina et al., 2018, 2020; Filazzola & Lortie, 2022; Gomes et al., 2022; Ivimey-Cook et al., 2025; Maitner et al., 2024; Pick et al., 2025; Purgar et al., 2022; Sánchez-Tójar et al., 2025) and workflows (Blischak et al., 2016; Braga et al., 2023; Brousil et al., 2023; Erickson et al., 2024; Hurley et al., 2022; Kim et al., 2022; Sequeira et al., 2021; Stoudt et al., 2021; E. P. White et al., 2019; Yenni et al., 2019).

Throughout this thesis I use the term 'reproducibility' in a broad sense, like in the phrase 'reproducibility crisis,' encapsulating both *replicability* and *(computational) reproducibility*. Following Fraser et al. (2020), replicability refers to the ability to repeat a study and obtain consistent results. *Direct* replication involves repeating the entire study, including data collection and analysis, as closely as possible to evaluate the validity of the original findings, while *conceptual* replications make deliberate variations in the original study to test the generalisability and robustness of the original findings. *(Computational) reproducibility* is a related but distinct concept which focuses on computational aspects of a study, in which the same data is re-analysed using the original analysis scripts (direct) or with different analytical methods (conceptual). Despite the attention reproducibility concerns have received across diverse disciplines and methodologies (e.g. Baldwin et al., 2022; Cockburn et al., 2020; Evans et al., 2023; Semmelrock et al., 2023; Thibault et al., 2023), the challenges specific to modelling remain under-explored. This thesis addresses that gap by focusing specifically on the reproducibility issues that plague ecological modelling. It is fundamentally concerned with *model-based research*, by which I mean: any research study that uses *quantitative modelling* to answer its research question, whether that be basic or applied in nature (Brudvig, 2017; Carmel et al., 2013; Houlahan et al., 2016), where a *quantitative model* is any model that uses mathematical or statistical concepts to describe or forecast how a system behaves, with parameters and their relationships expressed as numerical quantities (García-Díaz et al., 2019).

Models inform ecological understanding and theory, which guide the design or framing of experimental and observational studies and the hypotheses they are designed to test (Brudvig, 2017). Knowledge and evidence generated from empirical research on observational and experimental data (the "primary evidence base") is collated and synthesised to describe, understand and forecast ecological structure, patterns and processes in the system of interest (Betts et al., 2021; Houlahan et al., 2016; Jakeman et al., 2008; Letcher et al., 2013).Models may be used either directly to inform management decisions by synthesising available knowledge systematically and transparently, or indirectly through formal decision-support systems that facilitate the exploration and resolution of uncertainty about the efficacy of alternative management interventions (Addison et al., 2013; Dicks et al., 2014; García-Díaz et al., 2019; Gardner et al., 2018).

Many ecologists and conservation scientists work to collect data and provide tools that support improved decisions. Decision makers can draw on the primary-evidence base or

else directly on decision-support tools to inform their decision-making. Thus, ecological decision-making hinges on the reliability of ecological models, irrespective of whether they are developed in basic or applied research contexts (N. J. Bennett et al., 2017; Connolly et al., 2017; DeAngelis et al., 2021; Eker et al., 2018; García-Díaz et al., 2019). Many environmental decisions tackle "wicked problems" shaped by compounding anthropogenic threats in complex socio-economic and political contexts, with multiple stakeholders and multiple and/or competing objectives (Conroy & Peterson, 2013; Geary et al., 2020; Hemming et al., 2021; Van Der Sluijs, 2002). Environmental decisions may have irreversible consequences for species, ecosystems and society; but many of these problems require rapid decisions based upon available but uncertain evidence. It is therefore critical that ecological models are transparent to support trust in the scientific process and subsequent decisions (Bodner et al., 2021; Burgman et al., 2023; García-Díaz et al., 2019; Gomes, 2025; Powers & Hampton, 2018; Reichman et al., 2011; Schuwirth et al., 2019; Shoemaker & Loope, 2025; Travers et al., 2019). The reproducibility of ecological models is not merely an academic concern but is fundamental to the effectiveness of environmental decision-making.

1.1 Reproducibility & Transparency in Ecological Modelling

A pervasive problem in ecological research is the lack of transparency in how analytical decisions are made and reported. While transparency problems might initially appear to be primarily about communication and reporting, they are fundamentally connected to deeper issues of research reproducibility and reliability. When analytical decisions remain hidden, it becomes impossible to distinguish between rigorous research in which analysis decisions are carefully reasoned and justified, versus research that may be unconsciously or deliberately biased toward particular outcomes, or to identify when findings are contingent upon a specific set of modelling choices. This connection between transparency and reproducibility becomes particularly problematic when we consider the vast number of analytical choices, or "researcher degrees of freedom," (Simmons et al., 2011) available to modellers.

Research transparency is the practice of making all aspects of the scientific process explicit, clear and accessible, including theoretical, methodological and analytical decisions made during research (Evans et al., 2023). Lupia & Elman (2014) usefully distinguish between *production transparency* – where research artefacts like data and materials are shared and made open-access, including the data collection procedures generating them – and *analytic*

transparency, wherein a complete account of how analytic conclusions are drawn from the data is provided, and all steps in the modelling process, including assumptions and expectations about model outputs are clearly communicated (Bodner et al., 2020).

Concerns about the state of model reporting in ecological decision-support have been raised repeatedly, with many reviews arguing that current practices are neither transparent nor comprehensive (de Vos et al., 2011; Schmolke et al., 2010). Despite numerous attempts to promote thorough and transparent model documentation, particularly within the 'good modelling practice' literature (Alexandrov et al., 2011; Jakeman et al., 2006), assessments from over a decade ago underscoring the limited extent of model documentation and reporting (e.g. de Vos et al., 2011) still hold true today. While efforts to improve research transparency in ecology have increased over the last decade, they have primarily focused on *production transparency*, with growing emphasis on data and code sharing (O'Dea, Parker, et al., 2021). However, many obstacles persist, and interventions to improve transparency and openness still do not adequately ensure *analytic transparency* or research reproducibility (O'Dea, Parker, et al., 2021; Zvereva & Kozlov, 2021). Notably, while the open science movement in ecology has paid little attention to modelling specifically, the 'good modelling practice' literature represents a parallel effort within ecology that, although not yet fully integrated with the broader reproducibility movement, has consistently documented and argued for the centrality of transparency to good modelling practice. This context sets the stage for a more detailed examination of current transparency practices in ecological modelling and the challenges that remain.

Transparency in ecological modelling is routinely compromised because modelling documentation, including critical elements like data, code, metadata, methods and outputs, are often incomplete, fragmented across multiple publications, or ambiguous (de Vos et al., 2011). Similar to ecology more broadly, where code sharing rates range from 6-27% while data sharing reaches up to 79% (Čulina et al., 2020; Maitner et al., 2024), code sharing lags behind data sharing in ecological modelling, with consistently low rates of functional code sharing compared to data sharing (28% - 93% Feng et al., 2019; Kellner et al., 2024). Agent-based models demonstrate particularly poor transparency with only ~11% sharing code (Janssen et al., 2020), while connectivity modelling studies show even lower rates with only 2% providing all required code and 13% making all raw data available, with only two studies providing both (Riordan-Short et al., 2023). Deep learning models (Ahmed et al., 2025), Species Distribution Models (SDMs, Feng et al., 2019) and simulation studies

(Williams et al., 2024) have relatively higher code sharing rates (30%, <50% and 68%, respectively). However, the quality and completeness of shared code is a major barrier to computational reproducibility (Kellner et al., 2024; Williams et al., 2024).

Metadata and documentation of computational workflows that systematically links data, models and results, are frequently lacking, making data unusable and limiting reproduction (Barros et al., 2023; Essawy et al., 2020; Koivisto & Mäntylä, 2024; McIntire et al., 2022). Where documentation does exist, it is not standardised, often incomplete, and may only be understandable to the original author (Ayllón et al., 2021). Model descriptions are incompletely documented, with equations being reported for only some components or sub-models (de Vos et al., 2011). Similarly, critical parameters and hyperparameters are not disclosed in most published models (Ahmed et al., 2025; Feng et al., 2019; Morrison et al., 2016), and when they are, the published manuscript version does not always match the value or equation recorded in the code (de Vos et al., 2011).

Transparency in descriptions of final models has improved as adherence to standardised model reporting guidelines has increased, e.g. for agent-based modelling (Janssen et al., 2020). However, the modelling process itself, including underlying assumptions and rationale for methodological choices, remains opaque. Crucial assumptions that underpin the conceptualisation and implementation of the model are frequently underreported, including the model purpose and objectives, and biological rationales underlying choices (Arnqvist, 2020; Jakeman et al., 2006). Similarly, reporting on how parameter values are chosen, algorithms applied, or models calibrated is inconsistent and often incomplete (Feng et al., 2019; Kass et al., 2021; Szetey et al., 2025; Terry et al., 2024). The presentation of only minimal rather than full models from model simplification and formal model selection procedures is widespread, hiding model structural uncertainty and opportunities for multiple hypothesis testing (Forstmeier & Schielzeth, 2011; E. G. Simmonds et al., 2024).

1.1.1 Causes of Unsatisfactory Model Reporting Practices

Multiple technical, structural and cultural barriers perpetuate a culture of poor reporting practice in ecological modelling (Grimm et al., 2014). These barriers are intersecting and widespread within an academic research culture that rewards novelty, high publication

output, and particular types of findings over rigour, transparency, replication and reproducibility (Buxton et al., 2021; Josefsson et al., 2020; Nakagawa et al., 2025; Nissen et al., 2016; O'Dea, Parker, et al., 2021; Ottaviani et al., 2025; Seppelt et al., 2018). Implementing transparent and reproducible modelling is time-consuming and labour-intensive due to the size and complexity of some ecological models (de Vos et al., 2011; Essawy et al., 2020); an investment of effort that is often underestimated by funding agencies and not explicitly accounted for in project timelines or budgets (Micheletti et al., 2024). Moreover, open code, data, workflows and model documentation receive little formal recognition or reward compared to publications when it comes to career advancement (Borregaard & Hart, 2016; Koivisto & Mäntylä, 2024; Micheletti et al., 2024; Westoby et al., 2021). Short-term funding cycles and the pressure to produce recognised deliverables disincentivise investing the necessary time in learning new skills or meticulously documenting all aspects of the model and modelling process (Beck et al., 2020; Schuwirth et al., 2019).

Full transparency of ecological modelling requires additional metadata describing computational resources and linkages among data, models and results (Barros et al., 2023; Essawy et al., 2020). Ideally, this is aided by the use of workflow-management software (Brousil et al., 2023). Some predictive modelling approaches, like the PERFICT modelling framework (Barros et al., 2023; McIntire et al., 2022), are beginning to integrate all parts of the modelling cycle – development, parameterisation, execution, validation, and visualisation – in a transparent and reproducible workflow. However, whilst such approaches provide technical solutions and guidelines for promoting transparency, they are not yet the norm. Widespread adoption likely requires institutional and structural mechanisms that incentivise and reward modellers for following best practices in transparently documenting modelling workflows (Ivimey-Cook et al., 2025; O'Dea, Parker, et al., 2021; Sánchez-Tójar et al., 2025).

Although ecology has seen improvements in transparency due to an increasing number of journals now requiring open data and code (e.g. Freckleton, 2018; Thrall et al., 2023) there are not yet any equivalent policies, checklists or guidelines that stipulate the minimum details describing model rationale, structure or development, leading to substantial variation and omission in reporting across studies, as described above. Uptake of best-practices is limited by a lack of discipline-specific agreed upon standards and sufficient training about how models, model results and associated uncertainty should be reported (Davis & Kay, 2023; Koivisto & Mäntylä, 2024; Poisot et al., 2019; E. G. Simmonds et al., 2024),

perhaps because the modelling literature is plagued by inconsistent and contradictory use of terminology describing the modelling process (Schmolke et al., 2010). It is encouraging to see the development of application-specific reporting guidelines and checklists for ecological modelling (e.g. Achter et al., 2022; Feng et al., 2019; Fitzpatrick et al., 2021; Grimm et al., 2014, 2020; Planque et al., 2022; Riordan-Short et al., 2023; E. G. Simmonds et al., 2024; M. B. Simmonds et al., 2022; Williams et al., 2024; Zurell et al., 2020), but journal policy will be required to ensure adherence to these guidelines.

1.1.2 Consequences of Unsatisfactory Model Reporting Practices

Poor transparency in ecological modelling has serious cascading effects on both decision-making and science. Incomplete and opaque reporting of the model and modelling process risk the model being used in inappropriate applications or decision contexts (Grimm et al., 2014) because the underlying modelling choices, assumptions and quantitative uncertainty are not communicated (de Vos et al., 2011; E. G. Simmonds et al., 2024). Poorly documented models or documentation that is overly technical contribute to perceptions of models as 'black-boxes' among decision-makers and risk models being rejected by decision-makers (Ayllón et al., 2021; Eker et al., 2018; Herman et al., 2019; Roche et al., 2021; Weiskopf et al., 2022) for use in formal decision-support, instead turning to unstructured and subjective judgements (Addison et al., 2013; Grimm et al., 2014).

A lack of transparency can mask poor model design or serious flaws in the model, risking suboptimal decisions with negative or irreversible outcomes (Grimm et al., 2014; Regan et al., 2005; Schmolke et al., 2010). Uncertainty propagation stemming from seemingly small biases, omissions and miscalculations can generate large errors in predictions (Corkery et al., 2023). When negative ecological outcomes stem from flawed models, the credibility of models for use in decision-support risks being undermined, contributing further to the "knowing-doing" gap and to research waste (Buxton et al., 2021; Gardner et al., 2018; Hulme, 2014; A. T. Knight et al., 2008; Nakagawa et al., 2025; Purgar et al., 2022).

Poor modelling transparency hampers effective peer-review because the information required for thorough model evaluation is incomplete or absent (Jakeman et al., 2006; Schmolke et al., 2010). Consequently, models are frequently assessed primarily on scientific novelty, shifting the burden of credibility evaluation to readers, especially for complex models (Alexandrov et al., 2011; Schmolke et al., 2010). Model complexity is vaguely

defined in the literature and exists on a spectrum. Generally speaking, 'complex' models tend to have a larger number of parameters, and greater structural complexity; they might incorporate non-linear processes with complex dependency structures; or they may employ computationally intensive algorithms (McCrea et al., 2023; Naas et al., 2024). Complex models may require complicated or additional analyses to ensure they are interpretable (Lucas et al. 2020), or may require additional steps to construct, validate and evaluate them when compared to simpler models. They are more likely (but not always) to be mechanistic, as opposed to correlative, and predictive rather than explanatory.Without adequate model evaluation it is impossible to evaluate the model's fitness for purpose and to calibrate confidence in inferences drawn from them (Rykiel Jr., 1996). Poor transparency therefore makes it impossible to evaluate the reliability of evidence (O'Dea, Parker, et al., 2021), obscuring errors and preventing reanalysis and replication.

Poorly documented code, data and models inhibit reproducibility and make reanalysis and replication difficult (Bylemans et al., 2025), thereby preventing verification of results and findings. This lack of standardisation and transparency extends to poor reporting of modelling choices and performance metrics, which makes synthesis, interpretation and comparison across studies difficult (Christie et al., 2021; Desbureaux, 2021). Synthesis of multiple datasets is critical for answering global or macro ecological scale questions, for meta-analysis, and for monitoring, where datasets are combined through time as new data arrives (Beck et al., 2020; Yenni et al., 2019). Scientific progress can be hampered as researchers are unable to build directly on prior work (Čulina et al., 2018; Powers & Hampton, 2018). Finally, poor transparency provides opportunities for undisclosed researcher degrees of freedom and questionable research practices (QRPs), which significantly threaten reproducibility.

1.2 Researcher Degrees of Freedom: A Hidden Driver of Irreproducibility

'Researcher degrees of freedom,' or the many decisions and alternative analysis choices made throughout the research cycle arise due to analytic flexibility (Simmons et al., 2011; Wicherts et al., 2016) and create what Gelman & Loken (2013) termed the "garden of forking paths" – a vast space of alternative analytical approaches, any of which might seem reasonable in isolation. In ecological modelling studies, where modellers are faced with determining the

appropriate level of simplification required to produce useful model outputs while capturing important ecosystem dynamics and patterns,the number of defensible analytical pathways can multiply combinatorially, encompassing decisions about data preprocessing, how to operationalise and represent pattern and process, variable selection, model specification, model evaluation, application and interpretation of results (Allison et al., 2018; T. Arnold et al., 2020; Engelschalt et al., 2023; Grimm & Berger, 2016; Jakeman et al., 2006; Lahtinen et al., 2017; Liu et al., 2020; Poliseli et al., 2022; Upmeier zu Belzen et al., 2021). Emerging metaresearch from psychology and other fields has evidenced that analytical choices can substantially influence research conclusions (Botvinik-Nezer et al., 2020; Schweinsberg et al., 2021; Silberzahn et al., 2018), undermining reproducibility (see also Chapter 2 for a brief review). However, there is limited empirical evidence about the degree to which published findings are contingent on analytic choices in ecology, despite long-held theoretical concerns (e.g. Gelman & Loken, 2013), practical concerns (e.g. Giljohann et al., 2015), a methodological tradition of approaches aiming to address them (e.g. sensitivity and robustness analyses, model averaging and ensemble modelling, Saltelli et al., 2019; Conn et al., 2018; Dormann et al., 2018; Grimm & Berger, 2016) and an emerging literature investigating the influence of analysis choices on results in response to concerns about the reproducibility crisis (e.g. Bylemans et al., 2025; Commander et al., 2022; Desbureaux, 2021; Marshall & Duthie, 2024).

The combination of extensive analytical choices and a culture of poor transparency create the conditions for a perfect storm of reproducibility problems in ecological modelling. Researcher degrees of freedom are not inherently problematic but undermine reproducibility (beyond the computational) when they are used opportunistically or systematically without disclosure (Bakker et al., 2020; Wicherts et al., 2016). These behaviours are collectively termed 'questionable research practices' (QRPs, Neoh et al., 2023) and include practices like *p*-hacking, hypothesising after results are known (HARKing) and selective reporting of results (See Table 4.1 in Chapter 4 for a summary of QRPs in ecology and evolutionary biology). Importantly QRPs are distinct from deliberate fraud (Wigboldus & Dotsch, 2015). Publication bias and a "publish or perish" research culture incentivises analytic strategies that maximise publication success rather than methodological rigour, and within this environment, unconscious biases lead researchers to choose analytical approaches that yield "publishable" results without explicit awareness (Agnoli et al., 2017; Gelman & Loken, 2013; Hindman, 2015).

The impacts of QRPs have been well-characterised in null hypothesis significance testing (NHST) research, where they bias the literature by increasing the probability of false-positives (Earp & Trafimow, 2015; Seaman & Weber, 2015) and the magnitude of effect-size estimates (Parker, Forstmeier, et al., 2016; Wicherts et al., 2016). However, their impacts on other types of research common in ecology, such as ecological modelling, is not yet understood. In NHST research, publication bias selects for significant findings with larger effect sizes over null results (Ioannidis et al., 2014; Parker, Forstmeier, et al., 2016; Yang et al., 2023). Whereas publication bias leaves null effects unsubmitted in the 'file drawer' (Lortie et al., 2007; McElreath & Smaldino, 2015), QRPs effectively move unpublishable research out of the file drawer and into the domain of published literature, further distorting the literature.

Current frameworks for understanding QRPs are NHST focused and ill-suited to ecological modelling, which poses unique methodological challenges that differentiate it from disciplines where reproducibility problems have been extensively studied. Unlike NHST and controlled experimental designs undertaken in fields like experimental psychology or clinical medicine, ecological modelling approaches applied to observational data involve iterative, adaptive analytical processes where data dependent decisions are not only common but often methodologically necessary (Grimm & Berger, 2016; Hämäläinen & Lahtinen, 2016). Model construction and development involves iterative refinement of the models whereby model complexity is built incrementally based on diagnostic results, and analytical approaches are commonly tailored to unique characteristics of specific datasets or research contexts – often this involves adapting to the results of model checking, for example, discovering distributional violations may require changing model specifications. Existing frameworks for understanding QRPs treat these data-dependent decisions as inherently problematic, yet they are integral to sound ecological modelling practice.

Similarly, existing approaches to addressing transparency and reproducibility problems need to be adapted for the ecological modelling context. For instance, methodological reforms like preregistration and registered reports have become important tools for mitigating threats to reproducibility from researcher degrees of freedom in some scientific disciplines and are even mainstream in leading journals like *Psychological Science* (Nosek et al., 2018; van den Akker et al., 2024). However, these tools assume linear NHST research designs, and need to be tailored to allow for iterative, adaptive analytical processes that are an essential component of ecological modelling (Chapter 5). Unless this problem is resolved,

the mismatch between existing solutions and the practicalities of ecological research will continue to result in low uptake of open science practices in ecology, beyond data and code sharing.

The preceding discussion highlights two conceptually distinct sources of problematic practice in ecological modelling that are worth distinguishing explicitly, as they require different remedies. The first issue is the primary focus of this thesis: the opportunistic misuse of researcher degrees of freedom driven by incentive structures that select for the appearance of model credibility over methodological rigour. These are precisely the conditions under which QRPs emerge, as discussed in Section 1.2. The second issue is poor practice arising from insufficient formal training in modelling, model validation, uncertainty quantification, and reproducible workflows. This is a longstanding concern documented extensively in the Good Modelling Practice (GMP) literature (e.g., Jakeman et al., 2006; Grimm et al., 2014; Jakeman et al., 2024; Micheletti et al., 2024).

Although these two problems may produce superficially similar outcomes in terms of modelling practice, their causal origins differ: the former is driven by incentives and norms, while the latter reflects insufficient quantitative training and mentorship. Consequently, they require different solutions; preregistration and open-science interventions are intended to target the former and are not substitutes for adequate methodological training. In practice, however, the two problems intersect; for example, if training normalises questionable research practices. Notably, many of the problems the GMP literature has long sought to address – for example, incomplete validation, opaque model documentation, and insufficient uncertainty reporting - substantially overlap with the transparency and reproducibility concerns motivating this thesis, suggesting that these two bodies of work have much to gain from closer dialogue.

1.3 Problem Statement & Thesis Aims

My purpose is to bring renewed attention to reproducibility in ecology, specifically ecological modelling. As others have observed, ecology is at risk of a reproducibility crisis (Fidler et al., 2017), with the credibility and reliability of model-based research threatened by intersecting issues of transparency and researcher degrees of freedom. Although ecology has seen substantial improvements in data and code sharing, it is yet to reckon with the 'flexibility' in analytic decisions and any ensuing reproducibility issues. In this thesis, I focus

on *model-based research*. I have set the scope broadly because the methodological literature on modelling is siloed between disciplines and cultures of practice (Bodner et al., 2020; Saltelli et al., 2019). Although there may certainly be context-dependencies associated with particular strands of ecological modelling, many of the issues concerning reproducibility and transparency identified in this thesis are fundamental to most model-based research – I wish to avoid prematurely limiting the applicability of the findings in this thesis.

This thesis investigates the reproducibility and transparency of ecological modelling, with a particular focus on applied research. I examine the fundamental tension between maintaining methodological flexibility essential for complex modelling and the need to constrain researcher degrees of freedom to ensure reproducible, credible results. Each core chapter is centred around the concept of 'researcher degrees of freedom' and research transparency. Chapter 2 and Chapter 4 aim to characterise reproducibility and transparency issues in ecological modelling, demonstrating how researcher degrees of freedom can undermine reproducibility, particularly in the context of a culture of poor model reporting. Chapter 2 provides an empirical demonstration of the "garden of forking paths" highlighting the hidden ways that legitimate analytical choices can fundamentally alter ecological understanding. While Chapter 4 underscores the threat of QRPs in ecological modelling when researcher degrees of freedom remain undisclosed and opportunistically exploited. Chapter 5 develops adaptive solutions tailored to the discipline's specific methodological requirements that improve reproducibility and transparency without sacrificing necessary flexibility.

My research seeks to address an empirical gap quantifying the magnitude and consequences of analytic variability, a conceptual gap in understanding QRPs in model-based research, and an implementation gap in the availability and suitable of open science tools for mitigating the risks of researcher degrees of freedom to reproducibility and transparency in ecological modelling. Below, I outline my research questions intended to address these gaps in understanding and methodological reform, while summarising the work of my core chapters in answering those questions.

Chapters 2 and 3: The 'Same Data, Different Analysts' Study

While theoretical concerns about researcher degrees of freedom are well-established, there is limited empirical evidence about the degree to which analytical decisions affect research outcomes in ecological contexts. In Chapter 2, I seek to answer, to what extent do different analytical approaches applied to the same datasets yield different conclusions, and what

factors explain this variability? This study is the first of its kind within ecology, implementing a large-scale 'many-analyst' methodology to quantify the degree and extent of analytic variability in ecology and evolutionary biology, crowdsourcing 174 teams of 246 researchers to investigate the same research question using two different datasets, one from each field. This chapter empirically quantifies analytic heterogeneity by conducting a meta-analysis of analysts' results and examines how analytical decisions affect research outcomes when researchers have identical data and research questions by investigating whether particular features of analyses – such as the peer-reviewed quality, uniqueness of variable selection and choice to use random effects in analysts' models – contributed to among-study differences.

Chapter 3 presents a brief software paper describing the packaged *R* code used to implement the analyses in Chapter 2. The *ManyEcoEvo* package serves three purposes, focused on different types of replication. In the first instance, it served as a tool for conducting the analysis in Chapter 3 reproducibly and transparently, thereby allowing others to computationally reproduce our study. Secondly, it facilitates direct replication by allowing others to replicate the 'Same Data, Different Analysts' study using new data, and conceptual replication so that others can replicate our study with different analytical decisions. Finally, it facilitates computational analysis for new many-analyst studies.

Chapter 4: A Framework for Questionable Research Practices in Ecological Modelling

Current frameworks for understanding and classifying QRPs are designed for NHST-research, and do not adequately address the complexities of model-based ecological research. We need suitable conceptual frameworks to understand the causes, types and consequences of QRPs so that we can identify relevant practices in model-based research to understand when, where and how they might occur. Only then can we characterise the prevalence and impacts of QRPs in modelling, and design appropriate solutions that protect against bias caused by QRPs. In Chapter 4, I ask how can we adequately conceptualise QRPs to distinguish between legitimate, methodologically driven analytic decisions and opportunistic, results-driven choices in the context of iterative ecological modelling?

Chapter 4 aims to address a critical gap in how QRPs are understood and defined by developing a conceptual framework for understanding how QRPs emerge specifically in ecological modelling contexts. I first review the literature to identify different classes of QRPs that can occur in ecological modelling, building a map of QRPs by locating examples of specific practices from each class within different stages of the modelling process. I

then extend Gelman & Loken's (2013) mathematical framework describing the "garden of forking paths" to modelling contexts. I include this mathematical formalisation to help distinguish defensible and opportunistic data-dependent modelling choices, which subsequently facilitates a definition of QRPs relevant to ecological modelling.

Chapter 5: Adaptive Preregistration for Model-Based Research

Effective solutions require understanding of the problem they aim to address, consequently in Chapter 5 I build on the findings from Chapters 2 and 4 to extend preregistration to model-based research in ecology and conservation. The critical barrier confronting this task was how to constrain researcher degrees of freedom sufficiently to mitigate QRPs, while permitting a certain degree of flexibility necessary for iterative model development and data-dependent analytic decisions? My answer to this question is '*Adaptive* Preregistration,' which expands upon the existing model of preregistration to accommodate these requirements through the mechanisms of 'registered flexibility' and 'interim preregistrations,' in tandem with pilot analyses and data partitioning when required. In this chapter I also present a comprehensive preregistration template for ecological modelling, reflecting common activities and decisions made throughout the modelling process. This chapter culminates in a case-study of environmental flows modelling for decision-support in Victoria, Australia, through which I qualitatively evaluate the preregistration template and Adaptive Preregistration methodology. The case study preregistration provides an example to guide others in completing their own. To help make the process of Adaptive Preregistration readily applicable, I have also provided a user guide to Adaptive Preregistration (as a GitHub webpage at egouldo.github.io/EcoConsPreReg, reproduced in Section D.3) that summarises its core principles and outlines a technical procedure for implementation, given that existing registries are not equipped to handle registered flexibility or interim preregistrations.

Chapter 6: General Discussion

Finally, in Chapter 6, I discuss the principal findings from each core chapter, inferring implications for research practice and recommendations for improving the reproducibility and transparency of ecological modelling. I draw to a close by highlighting the contributions my research makes towards broader efforts to improve scientific practice.

2 "Same Data, Different Analysts:" Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology

Preface

This chapter includes the peer-reviewed stage 2 registered report published in BMC Biology:

> **Gould, E.**, Fraser, H. S., Parker, T. H. et al. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. BMC Biology 23, 35 (2025). https://doi.org/10.1186/s12915-024-02101-x.

The stage one registered report received in-principle acceptance on March 11 2020, and was uploaded to the BMC Biology Registered Report FigShare collection:

> Parker, T., Fraser, H. S., Nakagawa, S., **Gould, E.**, Griffith, S., Vesk, P. A., & Fidler, F. (2020). Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology [Registered Report Stage 1 Protocol] (Version 1). figshare. https://doi.org/10.6084/m9.figshare.12034833.v1.

The content included in this chapter is identical to that reported in the peer-reviewed publication, however, the formatting has been altered to match the style of the thesis. Some figures have been reformatted for readability. The published article is attached in Appendix B, reproduced with permission under license CC BY-NC-ND. All online supplementary material is presented in Appendix A, excluding code, and has also been reformatted in the style of the thesis. Note that Appendices 1 and 2 of the published article have been incorporated under Appendix A.5 in Table A.20 and Table A.21, respectively.

The manuscript is written as a completely reproducible quarto document (Gould et al., 2024), hosted publicly on GitHub Pages at egouldo.github.io/ManyAnalysts. Code for all results reported in this Chapter are described in Gould et al. (2024) and the *ManyEcoEvo* R package used to conduct all reported analyses is described in Chapter 3.

Abstract

Although variation in effect sizes and predicted values among studies of similar phenomena is inevitable, such variation far exceeds what might be produced by sampling error alone. One possible explanation for variation among results is differences among researchers in the decisions they make regarding statistical analyses. A growing array of studies has explored this analytical variability in different fields and has found substantial variability among results despite analysts having the same data and research question. Many of these studies have been in the social sciences, but one small "many analyst" study found similar variability in ecology. We expanded the scope of this prior work by implementing a large-scale empirical exploration of the variation in effect sizes and model predictions generated by the analytical decisions of different researchers in ecology and evolutionary biology. We used two unpublished datasets, one from evolutionary ecology (blue tit, *Cyanistes caeruleus*, to compare sibling number and nestling growth) and one from conservation ecology (*Eucalyptus*, to compare grass cover and tree seedling recruitment). The project leaders recruited 174 analyst teams, comprising 246 analysts, to investigate the answers to prespecified research questions. Analyses conducted by these teams yielded 141 usable effects (compatible with our meta-analyses and with all necessary information provided) for the blue tit dataset, and 85 usable effects for the *Eucalyptus* dataset. We found substantial heterogeneity among results for both datasets, although the patterns of variation differed between them. For the blue tit analyses, the average effect was convincingly negative, with less growth for nestlings living with more siblings, but there was near continuous variation in effect size from large negative effects to effects near zero, and even effects crossing the traditional threshold of statistical significance in the opposite direction. In contrast, the average relationship between grass cover and *Eucalyptus* seedling number was only slightly negative and not convincingly different from zero, and most effects ranged from weakly negative to weakly positive, with about a third of effects crossing the traditional threshold of significance in one direction or the other. However, there were also several striking outliers in the *Eucalyptus* dataset, with effects far from zero. For both datasets, we

found substantial variation in the variable selection and random effects structures among analyses, as well as in the ratings of the analytical methods by peer reviewers, but we found no strong relationship between any of these and deviation from the meta-analytic mean. In other words, analyses with results that were far from the mean were no more or less likely to have dissimilar variable sets, use random effects in their models, or receive poor peer reviews than those analyses that found results that were close to the mean. The existence of substantial variability among analysis outcomes raises important questions about how ecologists and evolutionary biologists should interpret published results, and how they should conduct analyses in the future.

Keywords: analytical heterogeneity; metascience; many-analyst; replication crisis; reproducibility.

2.1 Introduction

One value of science derives from its production of replicable, and thus reliable, results. When we repeat a study using the original methods we should be able to expect a similar result. However, perfect replicability is not a reasonable goal. Effect sizes will vary, and even reverse in sign, by chance alone (Gelman & Weakliem, 2009). Observed patterns can differ for other reasons as well. It could be that we do not sufficiently understand the conditions that led to the original result so when we seek to replicate it, the conditions differ due to some ‘hidden moderator’. This hidden moderator hypothesis is described by meta-analysts in ecology and evolutionary biology as ‘true biological heterogeneity’ (Senior et al., 2016). This idea of true heterogeneity is popular in ecology and evolutionary biology, and there are good reasons to expect it in the complex systems in which we work (Ellison et al., 2018). However, despite similar expectations in psychology, recent evidence in that discipline contradicts the hypothesis that moderators are common obstacles to replicability, as variability in results in a large ‘many labs’ collaboration was mostly unrelated to commonly hypothesised moderators such as the conditions under which the studies were administered (Klein et al., 2018). Another possible explanation for variation in effect sizes is that researchers often present biased samples of results, thus reducing the likelihood that later studies will produce similar effect sizes (Forstmeier et al., 2017; Fraser et al., 2018; Open Science Collaboration, 2015; Parker, Forstmeier, et al., 2016; Parker & Yang, 2023). It also may be that although researchers did successfully replicate the conditions,

the experiment, and measured variables, analytical decisions differed sufficiently among studies to create divergent results (Silberzahn et al., 2018; Simonsohn et al., 2015).

Analytical decisions vary among studies because researchers have many options. Researchers need to decide how to exclude possibly anomalous or unreliable data, how to construct variables, which variables to include in their models, and which statistical methods to use. Depending on the dataset, this short list of choices could encompass thousands or millions of possible alternative specifications (Simonsohn et al., 2015). However, researchers making these decisions presumably do so with the goal of doing the best possible analysis, or at least the best analysis within their current skill set. Thus it seems likely that some specification options are more probable than others, possibly because they have previously been shown (or claimed) to be better, or because they are more well known. Of course, some of these different analyses (maybe many of them) may be equally valid alternatives. Regardless, on probably any topic in ecology and evolutionary biology, we can encounter differences in choices of data analysis. The extent of these differences in analyses and the degree to which these differences influence the outcomes of analyses and therefore studies' conclusions are important empirical questions. These questions are especially important given that many papers draw conclusions after applying a single method, or even a single statistical model, to analyze a dataset.

The possibility that different analytical choices could lead to different outcomes has long been recognised (Gelman & Loken, 2013), and various efforts to address this possibility have been pursued in the literature. For instance, one common method in ecology and evolutionary biology involves creating a set of candidate models, each consisting of a different (though often similar) set of predictor variables, and then, for the predictor variable of interest, averaging the slope across all models (i.e. model averaging) (Burnham & Anderson, 2002; Grueber et al., 2011). This method reduces the chance that a conclusion is contingent upon a single model specification, though use and interpretation of this method is not without challenges (Grueber et al., 2011). Further, the models compared to each other typically differ only in the inclusion or exclusion of certain predictor variables and not in other important ways, such as methods of parameter estimation. More explicit examination of outcomes of differences in model structure, model type, data exclusion, or other analytical choices can be implemented through sensitivity analyses (e.g., Noble et al., 2017). Sensitivity analyses, however, are typically rather narrow in scope, and are designed to assess the sensitivity of analytical outcomes to a particular analytical choice rather than

to a large universe of choices. Recently, however, analysts in the social sciences have proposed extremely thorough sensitivity analysis, including 'multiverse analysis' (Steegen et al., 2016) and the 'specification curve' (Simonsohn et al., 2015), as a means of increasing the reliability of results. With these methods, researchers identify relevant decision points encountered during analysis and conduct the analysis many times to incorporate many plausible decisions made at each of these points. The study's conclusions are then based on a broad set of the possible analyses and so allow the analyst to distinguish between robust conclusions and those that are highly contingent on particular model specifications. These are useful outcomes, but specifying a universe of possible modelling decisions is not a trivial undertaking. Further, the analyst's knowledge and biases will influence decisions about the boundaries of that universe, and so there will always be room for disagreement among analysts about what to include. Including more specifications is not necessarily better. Some analytical decisions are better justified than others, and including biologically implausible specifications may undermine this process. Regardless, these powerful methods have yet to be adopted, and even more limited forms of sensitivity analyses are not particularly widespread. Most studies publish a small set of analyses and so the existing literature does not provide much insight into the degree to which published results are contingent on analytical decisions.

Despite the potential major impacts of analytical decisions on variance in results, the outcomes of different individuals' data analysis choices have only recently begun to receive much empirical attention. The only formal exploration of this that we were aware of when we submitted our Stage 1 manuscript were (1) an analysis in social science that asked whether male professional football (soccer) players with darker skin tone were more likely to be issued red cards (ejection from the game for rule violation) than players with lighter skin tone (Silberzahn et al., 2018) and (2) an analysis in neuroimaging which evaluated nine separate hypotheses involving the neurological responses detected with fMRI in 108 participants divided between two treatments in a decision making task (Botvinik-Nezer et al., 2020). Several others have been published since (e.g., Huntington-Klein et al., 2021; Breznau et al., 2022; Coretta et al., 2023; Schweinsberg et al., 2021), and we recently learned of an earlier small study in ecology (Stanton-Geddes et al., 2014). In the red card study, 29 teams designed and implemented analyses of a dataset provided by the study coordinators (Silberzahn et al., 2018). Analyses were peer reviewed (results blind) by at least two other participating analysts; a level of scrutiny consistent with standard

pre-publication peer review. Among the final 29 analyses, odds-ratios varied from 0.89 to 2.93, meaning point estimates varied from having players with lighter skin tones receive more red cards (odds ratio < 1) to a strong effect of players with darker skin tones receiving more red cards (odds ratio > 1). Twenty of the 29 teams found a statistically-significant effect in the predicted direction of players with darker skin tones being issued more red cards. This degree of variation in peer-reviewed analyses from identical data is striking, but the generality of this finding has only just begun to be formally investigated (e.g., Huntington-Klein et al., 2021; Breznau et al., 2022; Coretta et al., 2023; Schweinsberg et al., 2021).

In the neuroimaging study, 70 teams evaluated each of the nine different hypotheses with the available fMRI data (Botvinik-Nezer et al., 2020). These 70 teams followed a divergent set of workflows that produced a wide range of results. The rate of reporting of statistically significant support for the nine hypotheses ranged from 21%to 84%, and for each hypothesis on average, 20% of research teams observed effects that differed substantially from the majority of other teams. Some of the variability in results among studies could be explained by analytical decisions such as choice of software package, smoothing function, and parametric versus non-parametric corrections for multiple comparisons. However, substantial variability among analyses remained unexplained, and presumably emerged from the many different decisions each analyst made in their long workflows. Such variability in results among analyses from this dataset and from the very different red-card dataset suggests that sensitivity of analytical outcome to analytical choices may characterize many distinct fields, as several more recent many-analyst studies also suggest (Breznau et al., 2022; Huntington-Klein et al., 2021; Schweinsberg et al., 2021).

To further develop the empirical understanding of the effects of analytical decisions on study outcomes, we chose to estimate the extent to which researchers' data analysis choices drive differences in effect sizes, model predictions, and qualitative conclusions in ecology and evolutionary biology. This is an important extension of the meta-research agenda of evaluating factors influencing replicability in ecology, evolutionary biology, and beyond (Fidler et al., 2017). To examine the effects of analytical decisions, we used two different datasets and recruited researchers to analyze one or the other of these datasets to answer a question we defined. The first question was "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" To answer this question, we provided a dataset that includes brood size manipulations from 332 broods

conducted over three years at Wytham Wood, UK. The second question was "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" For this question, analysts used a dataset that includes, among other variables, number of seedlings in different size classes, percentage cover of different life forms, tree canopy cover, and distance from canopy edge from 351 quadrats spread among 18 sites in Victoria, Australia.

We explored the impacts of data analysts' choices with descriptive statistics and with a series of tests to attempt to explain the variation among effect sizes and predicted values of the dependent variable produced by the different analysis teams for both datasets separately. To describe the variability, we present forest plots of the standardised effect sizes and predicted values produced by each of the analysis teams, estimate heterogeneity (both absolute, τ^2, and proportional, I^2) in effect size and predicted values among the results produced by these different teams, and calculate a similarity index that quantifies variability among the predictor variables selected for the different statistical models constructed by the different analysis teams. These descriptive statistics provide the first estimates of the extent to which explanatory statistical models and their outcomes in ecology and evolutionary biology vary based on the decisions of different data analysts. We then quantified the degree to which the variability in effect size and predicted values could be explained by (1) variation in the quality of analyses as rated by peer reviewers and (2) the similarity of the choices of predictor variables between individual analyses.

2.2 Methods

This project involved a series of steps (1-6) that began with identifying datasets for analyses and continued through recruiting independent groups of scientists to analyze the data, allowing the scientists to analyze the data as they saw fit, generating peer review ratings of the analyses (based on methods, not results), evaluating the variation in effects among the different analyses, and producing the final manuscript.

2.2.1 Step 1: Select Datasets

We used two previously unpublished datasets, one from evolutionary ecology and the other from ecology and conservation.

Evolutionary ecology

Our evolutionary ecology dataset is relevant to a sub-discipline of life-history research which focuses on identifying costs and trade-offs associated with different phenotypic conditions. These data were derived from a brood-size manipulation experiment imposed on wild birds nesting in boxes provided by researchers in an intensively studied population. Understanding how the growth of nestlings is influenced by the numbers of siblings in the nest can give researchers insights into factors such as the evolution of clutch size, determination of provisioning rates by parents, and optimal levels of sibling competition (DeKogel, 1997; Nicolaus et al., 2009; Royle et al., 1999; Vander Werf, 1992; Verhulst et al., 2006). Data analysts were provided this dataset and instructed to answer the following question: "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?"

Researchers conducted brood size manipulations and population monitoring of blue tits at Wytham Wood, a 380 ha woodland in Oxfordshire, U.K (1º 20'W, 51º 47'N). Researchers regularly checked approximately 1100 artificial nest boxes at the site and monitored the 330 to 450 blue tit pairs occupying those boxes in 2001-2003 during the experiment. Nearly all birds made only one breeding attempt during the April to June study period in a given year. At each blue tit nest, researchers recorded the date the first egg appeared, clutch size, and hatching date. For all chicks alive at age 14 days, researchers measured mass and tarsus length and fitted a uniquely numbered, British Trust for Ornithology (BTO) aluminium leg ring. Researchers attempted to capture all adults at their nests between day 6 and day 14 of the chick-rearing period. For these captured adults, researchers measured mass, tarsus length, and wing length and fitted a uniquely numbered BTO leg ring. During the 2001-2003 breeding seasons, researchers manipulated brood sizes using cross fostering. They matched broods for hatching date and brood size and moved chicks between these paired nests one or two days after hatching. They sought to either enlarge or reduce all manipulated broods by approximately one fourth. To control for effects of being moved, each reduced brood had a portion of its brood replaced by chicks from the paired increased brood, and vice versa. Net manipulations varied from plus or minus four chicks in broods of 12 to 16 to plus or minus one chick in broods of 4 or 5. Researchers left approximately one third of all broods unmanipulated. These unmanipulated broods were not selected systematically to match manipulated broods in clutch size or laying date. We have mass and tarsus length data from 3720 individual chicks divided among 167 experimentally enlarged broods, 165 experimentally reduced broods, and 120 unmanipulated broods. The full list

of variables included in the dataset is publicly available (osf.io/hdv8m), along with the data (osf.io/qjzby).

Additional Explanation

Shortly after beginning to recruit analysts, several analysts noted a small set of related errors in the blue tit dataset. We corrected the errors, replaced the dataset on our OSF site, and emailed the analysts on 19 April 2020 to instruct them to use the revised data. The email to analysts is available here (osf.io/4h53z). The errors are explained in that email.

Ecology and conservation

Our ecology and conservation dataset is relevant to a sub-discipline of conservation research which focuses on investigating how best to revegetate private land in agricultural landscapes. These data were collected on private land under the Bush Returns program, an incentive system where participants entered into a contract with the Goulburn Broken Catchment Management Authority and received annual payments if they executed predetermined restoration activities. This particular dataset is based on a passive regeneration initiative, where livestock grazing was removed from the property in the hopes that the *Eucalyptus* spp. overstorey would regenerate without active (and expensive) planting. Analyses of some related data have been published (Miles, 2008; Vesk et al., 2016) but those analyses do not address the question analysts answered in our study. Data analysts were provided this dataset and instructed to answer the following question: "How does grass cover influence *Eucalyptus* spp. seedling recruitment?".

Researchers conducted three rounds of surveys at 18 sites across the Goulburn Broken catchment in northern Victoria, Australia in winter and spring 2006 and autumn 2007. In each survey period, a different set of 15 x 15 m quadrats were randomly allocated across each site within 60 m of existing tree canopies. The number of quadrats at each site depended on the size of the site, ranging from four at smaller sites to 11 at larger sites. The total number of quadrats surveyed across all sites and seasons was 351. The number of *Eucalyptus* spp. seedlings was recorded in each quadrat along with information on the GPS location, aspect, tree canopy cover, distance to tree canopy, and position in the landscape. Ground layer plant species composition was recorded in three 0.5 x 0.5 m sub-quadrats within each quadrat. Subjective cover estimates of each species as well as bare ground,

litter, rock and moss/lichen/soil crusts were recorded. Subsequently, this was augmented with information about the precipitation and solar radiation at each GPS location. The full list of variables included in the dataset is publicly available (osf.io/r5gbn), along with the data (osf.io/qz5cu).

2.2.2 Step 2: Recruitment and Initial Survey of Analysts

The lead team (TP, HF, SN, EG, SG, PV, DH, FF) created a publicly available document providing a general description of the project (osf.io/mn5aj). The project was advertised at conferences, via Twitter, using mailing lists for ecological societies (including Ecolog, Evoldir, and lists for the Environmental Decisions Group, and Transparency in Ecology and Evolution), and via word of mouth. The target population was active ecology, conservation, or evolutionary biology researchers with a graduate degree (or currently studying for a graduate degree) in a relevant discipline. Researchers could choose to work independently or in a small team. For the sake of simplicity, we refer to these as 'analysis teams' though some comprised one individual. We aimed for a minimum of 12 analysis teams independently evaluating each dataset (see sample size justification below). We simultaneously recruited volunteers to peer review the analyses conducted by the other volunteers through the same channels. Our goal was to recruit a similar number of peer reviewers and analysts, and to ask each peer reviewer to review a minimum of four analyses. If we were unable to recruit at least half the number of reviewers as analysis teams, we planned to ask analysts to serve also as reviewers (after they had completed their analyses), but this was unnecessary. Therefore, no data analysts peer reviewed analyses of the dataset they had analyzed. All analysts and reviewers were offered the opportunity to share co-authorship on this manuscript and we planned to invite them to participate in the collaborative process of producing the final manuscript. All analysts signed [digitally] a consent (ethics) document (osf.io/xyp68) approved by the Whitman College Institutional Review Board prior to being allowed to participate.

> **Preregistration Deviation**
>
> Due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We identified our minimum number of analysts per dataset by considering the number of effects needed in a meta-analysis to generate an estimate of heterogeneity (τ^2) with a 95% confidence interval that does not encompass zero. This minimum sample size is invariant regardless of τ^2. This is because the same t-statistic value will be obtained by the same sample size regardless of variance (τ^2). We see this by first examining the formula for the standard error, SE for variance, (τ^2) or (SE(τ^2)) assuming normality in an underlying distribution of effect sizes (K. Knight, 2000):

$$\mathrm{SE}(\tau^2) = \sqrt{\frac{2\tau^4}{n-1}} \tag{2.1}$$

and then rearranging the above formula to show how the t-statistic is independent of τ^2, as seen below.

$$t = \frac{\tau^2}{SE(\tau^2)} = \sqrt{\frac{n-1}{2}} \tag{2.2}$$

We then find a minimum n = 12 according to this formula.

2.2.3 Step 3: Primary Data Analyses

Analysis teams registered and answered a demographic and expertise survey (osf.io/seqzy). We then provided them with the dataset of their choice and requested that they answer a specific research question. For the evolutionary ecology dataset that question was "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" and for the conservation ecology dataset it was "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" Once their analysis was complete, they answered a structured survey (osf.io/neyc7), providing analysis technique, explanations of their analytical choices, quantitative results, and a statement describing their conclusions. They also were asked to upload their analysis files (including the dataset as they formatted it for analysis and their analysis code [if applicable]) and a detailed journal-ready statistical methods section.

Additional Explanation

As is common in many studies in ecology and evolutionary biology, the datasets we provided contained many variables, and the research questions we provided could be addressed by our datasets in many different ways. For instance, volunteer analysts had to choose the dependent (response) variable and the independent variable, and make numerous other decisions about which variables and data to use and how to structure their model.

Preregistration Deviation

We originally planned to have analysts complete a single survey (osf.io/neyc7), but after we evaluated the results of that survey, we realised we would need a second survey (osf.io/8w3v5) to adequately collect the information we needed to evaluate heterogeneity of results (step 5). We provided a set of detailed instructions with the follow-up survey, and these instructions are publicly available and can be found within the following files (blue tit: osf.io/kr2g9, *Eucalyptus*: osf.io/dfvym).

2.2.4 Step 4: Peer Reviews of Analyses

At minimum, each analysis was evaluated by four different reviewers, and each volunteer peer reviewer was randomly assigned methods sections from at least four analyst teams (the exact number varied). Each peer reviewer registered and answered a demographic and expertise survey identical to that asked of the analysts, except we did not ask about 'team name' since reviewers did not work in teams. Reviewers evaluated the methods of each of their assigned analyses one at a time in a sequence determined by the project leaders. We systematically assigned the sequence so that, if possible, each analysis was allocated to each position in the sequence for at least one reviewer. For instance, if each reviewer were assigned four analyses to review, then each analysis would be the first analysis assigned to at least one reviewer, the second analysis assigned to another reviewer, the third analysis assigned to yet another reviewer, and the fourth analysis assigned to a fourth reviewer. Balancing the order in which reviewers saw the analyses controls for order effects, e.g. a reviewer might be less critical of the first methods section they read than the last.

The process for a single reviewer was as follows. First, the reviewer received a description of the methods of a single analysis. This included the narrative methods section, the analysis team's answers to our survey questions regarding their methods, including analysis code,

and the dataset. The reviewer was then asked, in an online survey (osf.io/4t36u), to rate that analysis on a scale of 0-100 based on this prompt: "Rate the overall appropriateness of this analysis to answer the research question (*one of the two research questions inserted here*) with the available data. To help you calibrate your rating, please consider the following guidelines:

- 100: A perfect analysis with no conceivable improvements from the reviewer
- 75: An imperfect analysis but the needed changes are unlikely to dramatically alter outcomes
- 50: A flawed analysis likely to produce either an unreliable estimate of the relationship or an over-precise estimate of uncertainty
- 25: A flawed analysis likely to produce an unreliable estimate of the relationship and an over-precise estimate of uncertainty
- 0: A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise estimate of uncertainty that places undue confidence in the incorrect estimate. *Please note that these values are meant to calibrate your ratings. We welcome ratings of any number between 0 and 100.

After providing this rating, the reviewer was presented with this prompt, in multiple-choice format: "Would the analytical methods presented produce an analysis that is (a) publishable as is, (b) publishable with minor revision, (c) publishable with major revision, (d) deeply flawed and unpublishable?" The reviewer was then provided with a series of text boxes and the following prompts: "Please explain your ratings of this analysis. Please evaluate the choice of statistical analysis type. Please evaluate the process of choosing variables for and structuring the statistical model. Please evaluate the suitability of the variables included in (or excluded from) the statistical model. Please evaluate the suitability of the structure of the statistical model. Please evaluate choices to exclude or not exclude subsets of the data. Please evaluate any choices to transform data (or, if there were no transformations, but you think there should have been, please discuss that choice)." After submitting this review, a methods section from a second analysis was then made available to the reviewer. This same sequence was followed until all analyses allocated to a given reviewer were provided and reviewed. After providing the final review, the reviewer was simultaneously provided with all four (or more) methods sections the reviewer had just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation. The invitation

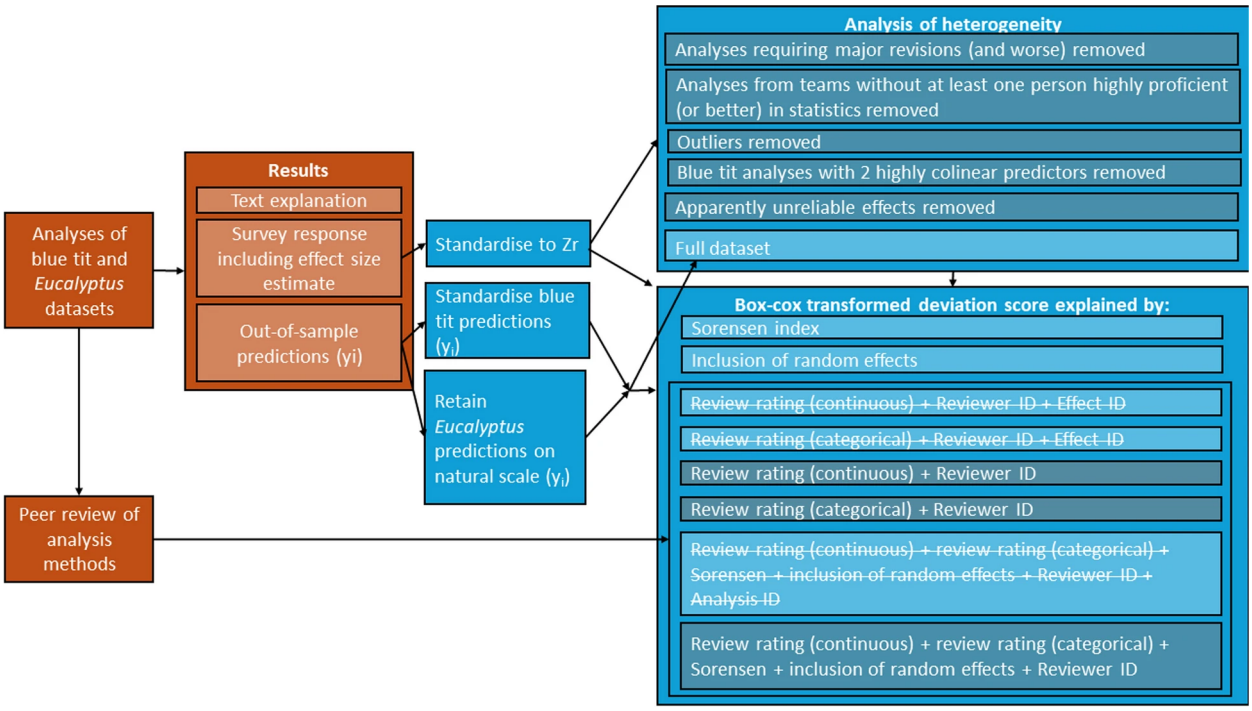


Figure 2.1: *Schematic of research process showing recruited analyst and reviewer contributions in orange and core team contributions in blue. Items that are crossed out were preregistered but could not be conducted. Items with a greyed background were added as exploratory analyses after preregistration.*

to revise the original ratings was as follows: "If, now that you have seen all the analyses you are reviewing, you wish to revise your ratings of any of these analyses, you may do so now." The text box was prefaced with this prompt: "Please explain your choice to revise (or not to revise) your ratings."

Additional Explanation: unregistered analysis

To determine how consistent peer reviewers were in their ratings, we assessed inter-rater reliability among reviewers for both the categorical and quantitative ratings combining blue tit and *Eucalyptus* data using Krippendorff's alpha for ordinal and continuous data respectively. This provides a value that is between -1 (total disagreement between reviewers) and 1 (total agreement between reviewers).

2.2.5 Step 5: Evaluate Variation

Additional Explanation: analysis schematic

The lead team conducted a range of preregistered and exploratory analyses to understand variation between analyses and their results. Figure 2.1 is intended to clarify the analyses described below.

The lead team conducted the analyses outlined in this section. We described the variation in model specification in several ways. We calculated summary statistics describing variation among analyses, including mean, SD, and range of number of variables per model included as fixed effects, the number of interaction terms, the number of random effects, and the mean, SD, and range of sample sizes. We also present the number of analyses in which each variable was included. We summarised the variability in standardised effect sizes and predicted values of dependent variables among the individual analyses using standard random effects meta-analytic techniques. First, we derived standardised effect sizes from each individual analysis. We did this for all linear models or generalised linear models by converting the t value and the degree of freedom (df) associated with regression coefficients (e.g. the effect of the number of siblings [predictor] on growth [response] or the effect of grass cover [predictor] on seedling recruitment [response]) to the correlation coefficient, r, using the following:

$$r = \sqrt{\frac{t^2}{(t^2 + df)}} \tag{2.3}$$

This formula can only be applied if t and df values originate from linear or generalised linear models (GLMs, Nakagawa & Cuthill, 2007). If, instead, linear mixed-effects models (LMMs) or generalised linear mixed-effects models (GLMMs) were used by a given analysis, the exact df cannot be estimated. However, adjusted df can be estimated, for example, using the Satterthwaite approximation of df, df_S, (note that SAS uses this approximation to obtain df for LMMs and GLMMs, Luke, 2017). For analyses using either LMMs or GLMMs that do not produce df_S we planned to obtain df_S by rerunning the same (G)LMMs using the `lmer()` or `glmer()` function in the *lmerTest* package in R (Kuznetsova et al., 2017; R Core Team, 2024).

Preregistration Deviation

Rather than re-run these analyses ourselves, we sent a follow-up survey (referenced above under “Primary data analyses”) to analysts and asked them to follow our instructions for producing this information. The instructions are publicly available and can be found within the following files (blue tit: osf.io/kr2g9, *Eucalyptus*: osf.io/dfvym).

We then used the t values and df_S from the models to obtain r as per the formula above. All r and accompanying df (or df_S) were converted to Fisher's Z_r

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{2.4}$$

and its sampling variance; $1/(n–3)$ where $n = df + 1$. Any analyses from which we could not derive a signed Z_r, for instance one with a quadratic function in which the slope changed sign, were considered unusable for analyses of Z_r. We expected such analyses would be rare. In fact, most submitted analyses excluded from our meta-analysis of Z_r were excluded because of a lack of sufficient information provided by the analyst team rather than due to the use of effects that could not be converted to Z_r. Regardless, as we describe below, we generated a second set of standardised effects (predicted values) that could (in principle) be derived from any explanatory model produced by these data.

Besides Z_r, which describes the strength of a relationship based on the amount of variation in a dependent variable explained by variation in an independent variable, we also examined differences in the shape of the relationship between the independent and dependent variables. To accomplish this, we derived a point estimate (out-of-sample predicted value) for the dependent variable of interest for each of three values of our primary independent variable. We originally described these three values as associated with the 25th percentile, median, and 75th percentile of the independent variable and any covariates.

Preregistration Deviation

The original description of the out-of-sample specifications did not account for the facts that (a) some variables are not distributed in a way that allowed division in percentiles and that (b) variables could be either positively or negatively correlated with the dependent variable. We provide a more thorough description here: We derived three point-estimates (out-of-sample predicted values) for the dependent variable of interest; one for each of three values of our primary independent variable that we specified. We also specified values for all other variables that could have been included as independent variables in analysts' models so that we could derive the predicted values from a fully specified version of any model produced by analysts. For all potential independent variables, we selected three values or categories. Of the

three we selected, one was associated with small, one with intermediate, and one with large values of one typical dependent variable (day 14 chick weight for the blue tit data and total number of seedlings for the *Eucalyptus* data; analysts could select other variables as their dependent variable, but the others typically correlated with the two identified here). For continuous variables, this means we identified the 25th percentile, median, and 75th percentile and, if the slope of the linear relationship between this variable and the typical dependent variable was positive, we left the quartiles ordered as is. If, instead, the slope was negative, we reversed the order of the independent variable quartiles so that the 'lower' quartile value was the one associated with the lower value for the dependent variable. In the case of categorical variables, we identified categories associated with the 25th percentile, median, and 75th percentile values of the typical dependent variable after averaging the values for each category. However, for some continuous and categorical predictors, we also made selections based on the principle of internal consistency between certain related variables, and we fixed a few categorical variables as identical across all three levels where doing so would simplify the modelling process (specification tables available: blue tit: osf.io/86akx; *Eucalyptus*: osf.io/jh7g5).

We used the 25th and 75th percentiles rather than minimum and maximum values to reduce the chance of occupying unrealistic parameter space. We planned to derive these predicted values from the model information provided by the individual analysts. All values (predictions) were first transformed to the original scale along with their standard errors (SE); we used the delta method (Ver Hoef, 2012) for the transformation of SE. We used the square of the SE associated with predicted values as the sampling variance in the meta-analyses described below, and we planned to analyze these predicted values in exactly the same ways as we analyzed Z_r in the following analyses.

Preregistration Deviation

1. Standardizing blue tit out-of-sample predictions y_i

Because analysts of blue tit data chose different dependent variables on different scales, after transforming out-of-sample values to the original scales, we standardised all values as z scores ('standard scores') to put all dependent variables on the same scale and make them comparable. This involved taking each relevant value on the original

scale (whether a predicted point estimate or a SE associated with that estimate) and subtracting the value in question from the mean value of that dependent variable derived from the full dataset and then dividing this difference by the standard deviation, SD, corresponding to the mean from the full dataset (Equation A.1). Thus, all our out-of-sample prediction values from the blue tit data are from a distribution with the mean of 0 and SD of 1.

Note that we were unable to standardise some analyst-constructed variables, so these analyses were excluded from the final out-of-sample estimates meta-analysis, see Appendix A.2.1 for details and explanation.

2. Log-transforming *Eucalyptus* out-of-sample predictions y_i

All analyses of the *Eucalyptus* data chose dependent variables that were on the same scale, that is, *Eucalyptus* seedling counts. Although analysts may have used different size-classes of *Eucalyptus* seedlings for their dependent variable, we considered these choices to be akin to subsetting, rather than as different response variables, since changing the size-class of the dependent variable ultimately results in observations being omitted or included. Consequently, we did not standardise *Eucalyptus* out-of-sample predictions.

We were unable to fit quasi-Poisson or Poisson meta-regressions, as desired (O'Hara & Kotze, 2010), because available meta-analysis packages (e.g. `metafor` and `metainc`) do not provide implementation for outcomes as estimates-only, methods are only provided for outcomes as ratios or rate-differences between two groups. Consequently, we log-transformed the out-of-sample predictions for the *Eucalyptus* data and use the mean estimate for each prediction scenario as the dependent variable in our meta-analysis with the associated SE as the sampling variance in the meta-analysis (Nakagawa, Ivimey-Cook, et al., 2023, Table 2).

We plotted individual effect size estimates (Z_r) and predicted values of the dependent variable (y_i) and their corresponding 95% confidence / credible intervals in forest plots to allow visualization of the range and precision of effect size and predicted values. Further, we included these estimates in random effects meta-analyses (Borenstein et al., 2017; Higgins et al., 2003) using the *metafor* package in R (Viechtbauer, 2010):

$$Z_r \sim 1 + (1|\text{analysisID}) \tag{2.5}$$

$$y_i \sim 1 + (1|\text{analysisID}) \tag{2.6}$$

where y_i is the predicted value for the dependent variable at the 25th percentile, median, or 75th percentile of the independent variables. The individual Z_r effect sizes were weighted with the inverse of sampling variance for Z_r. The individual predicted values for dependent variable (y_i) were weighted by the inverse of the associated SE^2 (original registration omitted "inverse of the" in error). These analyses provided an average Z_r score or an average y_i with corresponding 95% confidence interval and allowed us to estimate two heterogeneity indices, τ^2 and I^2. The former, τ^2, is the absolute measure of heterogeneity or the between-study variance (in our case, between-effect variance) whereas I^2 is a relative measure of heterogeneity. We obtained the estimate of relative heterogeneity (I^2) by dividing the between-effect variance by the sum of between-effect and within-effect variance (sampling error variance). I^2 is thus, in a standard meta-analysis, the proportion of variance that is due to heterogeneity as opposed to sampling error. When calculating I^2, within-study variance is amalgamated across studies to create a "typical" within-study variance which serves as the sampling error variance (Borenstein et al., 2017; Higgins et al., 2003). Our goal here was to visualize and quantify the degree of variation among analyses in effect size estimates (Nakagawa & Cuthill, 2007). We did not test for statistical significance.

Additional Explanation

Our use of I^2 to quantify heterogeneity violates an important assumption, but this violation does not invalidate our use of I^2 as a metric of how much heterogeneity can derive from analytical decisions. In standard meta-analysis, the statistic I^2 quantifies the proportion of variance that is greater than we would expect if differences among estimates were due to sampling error alone (Rosenberg, 2013). However, it is clear that this interpretation does not apply to our value of I^2 because I^2 assumes that each estimate is based on an independent sample (although these analyses can account for non-independence via hierarchical modelling), whereas all our effects were derived from largely or entirely overlapping subsets of the same dataset. Despite this, we

believe that I^2 remains a useful statistic for our purposes. This is because, in calculating I^2, we are still setting a benchmark of expected variation due to sampling error based on the variance associated with each separate effect size estimate, and we are assessing how much (if at all) the variability among our effect sizes exceeds what would be expected had our effect sizes been based on independent data. In other words, our estimates can tell us how much proportional heterogeneity is possible from analytical decisions alone when sample sizes (and therefore meta-analytic within-estimate variance) are similar to the ones in our analyses. Among other implications, our violation of the independent sample assumption means that we (dramatically) over-estimate the variance expected due to sampling error, and because I^2 is a proportional estimate, we thus underestimate the actual proportion of variance due to differences among analyses other than sampling error. However, correcting this underestimation would create a trivial value since we designed the study so that much of the variance would derive from analytic decisions as opposed to differences in sampled data. Instead, retaining the I^2 value as typically calculated provides a useful comparison to I^2 values from typical meta-analyses.

Interpretation of τ^2 also differs somewhat from traditional meta-analysis, and we discuss this further in the Results.

Finally, we assessed the extent to which deviations from the meta-analytic mean by individual effect sizes (Z_r) or the predicted values of the dependent variable (y_i) were explained by the peer rating of each analysis team's method section, by a measurement of the distinctiveness of the set of predictor variables included in each analysis, and by the choice of whether or not to include random effects in the model. The deviation score, which served as the dependent variable in these analyses, is the absolute value of the difference between the meta-analytic mean $\bar{Z}_r$ (or $\bar{y}_i$) and the individual Z_r (or y_i) estimate for each analysis. We used the Box-Cox transformation on the absolute values of deviation scores to achieve an approximately normal distribution (c.f. Fanelli & Ioannidis, 2013; Fanelli et al., 2017). We described variation in this dependent variable with both a series of univariate analyses and a multivariate analysis. All these analyses were general linear (mixed) models. These analyses were secondary to our estimation of variation in effect sizes described above. We wished to quantify relationships among variables, but we had no *a priori* expectation of effect size and made no dichotomous decisions about statistical significance.

Additional Explanation

In our meta-analyses based on Box-Cox transformed deviation scores, we leave these deviation scores unweighted. This is consistent with our registration, which did not mention weighting these scores. However, the fact that we did not mention weighting the scores was actually an error: we had intended to weight them, as is standard in meta-analysis, using the inverse variance of the Box-Cox transformed deviation scores Equation A.2. Unfortunately, when we did conduct the weighted analyses, they produced results in which some weighted estimates differed radically from the unweighted estimate because the weights were invalid. Such invalid weights can sometimes occur when the variance (upon which the weights depend) is partly a function of the effect size, as in our Box-Cox transformed deviation scores (Nakagawa et al., 2022). In the case of the *Eucalyptus* analyses, the most extreme outlier was weighted much more heavily (by close to two orders of magnitude) than any other effect sizes because the effect size was, itself, so high. Therefore, we made the decision to avoid weighting by inverse variance in all analyses of the Box-Cox transformed deviation scores. This was further justified because (a) most analyses have at least some moderately unreliable weights, and (b) the sample sizes were mostly very similar to each other across submitted analyses, and so meta-analytic weights are not particularly important here. We systematically investigated the impact of different weighting schemes and random effects on model convergence and results, see Appendix A.3.8 for more details.

When examining the extent to which reviewer ratings (on a scale from 0 to 100) explained deviation from the average effect (or predicted value), each analysis had been rated by multiple peer reviewers, so for each reviewer score to be included, we include each deviation score in the analysis multiple times. To account for the non-independence of multiple ratings of the same analysis, we planned to include analysis identity as a random effect in our general linear mixed model in the *lme4* package in R (Bates et al., 2015; R Core Team, 2024). To account for potential differences among reviewers in their scoring of analyses, we also planned to include reviewer identity as a random effect:

$$
\begin{aligned}
\text{DeviationScore}_j &= \text{BoxCox}(|\text{DeviationFromMean}_j|) \\
\text{DeviationScore}_{ij} &\sim \text{Rating}_{ij} + \\
&\quad \text{ReviewerID}_i + \\
&\quad \text{AnalysisID}_j \\
\text{ReviewerID}_i &\sim \mathcal{N}(0, \sigma_i^2) \\
\text{AnalysisID}_j &\sim \mathcal{N}(0, \sigma_j^2)
\end{aligned}
\tag{2.7}
$$

Where $\text{DeviationFromMean}_j$ is the deviation from the meta-analytic mean for the jth analysis, ReviewerID_i is the random intercept assigned to each i reviewer, and AnalysisID_j is the random intercept assigned to each j analysis, both of which are assumed to be normally distributed with a mean of 0 and a variance of σ^2. Absolute deviation scores were Box-Cox transformed using the `step_box_cox()` function from the *timetk* package in R (Dancho & Vaughan, 2023; R Core Team, 2024).

We conducted a similar analysis with the four categories of reviewer ratings ((1) deeply flawed and unpublishable, (2) publishable with major revision, (3) publishable with minor revision, (4) publishable as is) set as ordinal predictors numbered as shown here. As with the analyses above, we planned for these analyses to also include random effects of analysis identity and reviewer identity. Both of these analyses (1: 1-100 ratings as the fixed effect, 2: categorical ratings as the fixed effects) were planned to be conducted eight times for each dataset. Each of the four responses (Z_r, y_{25th}, y_{50th}, y_{75th}) were to be compared once to the initial ratings provided by the peer reviewers, and again based on the revised ratings provided by the peer reviewers.

Preregistration Deviation

1. We planned to include random effects of both analysis identity and reviewer identity in these models comparing reviewer ratings with deviation scores. However, after we received the analyses, we discovered that a subset of analyst teams had either conducted multiple analyses and/or identified multiple effects per analysis as answering the target question. We therefore faced an even more complex potential set of random effects. We decided that including Team ID and

Effect ID along with Reviewer ID as random effects in the same model would almost certainly lead to model fit problems, and so we started with simpler models including just Effect ID and Reviewer ID. However, even with this simpler structure, our dataset was sparse, with reviewers rating a small number of analyses, resulting in models with singular fit (Appendix A.3.2). Removing one of the random effects was necessary for the models to converge. For both models of deviation from the meta-analytic mean explained by categorical or continuous reviewer ratings, we removed the random effect of Effect ID, leaving Reviewer ID as the only random effect.

2. We conducted analyses only with the final peer ratings after the opportunity for revision, not with the initial ratings. This was because when we recorded the final ratings, the initial ratings were over-written, therefore we did not have access to those initial values.

The next set of univariate analyses sought to explain deviations from the mean effects based on a measure of the distinctiveness of the set of variables included in each analysis. As a 'distinctiveness' score, we used Sorensen's Similarity Index (an index typically used to compare species composition across sites), treating variables as species and individual analyses as sites. To generate an individual Sorensen's value for each analysis required calculating the pairwise Sorensen's value for all pairs of analyses (of the same dataset), and then taking the average across these Sorensen's values for each analysis. We calculated the Sorensen's index values using the *betapart* package (Baselga et al., 2023) in R:

$$\beta_{\text{Sorensen}} = \frac{b+c}{2a+b+c} \tag{2.8}$$

where a is the number of variables common to both analyses, b is the number of variables that occur in the first analysis but not in the second and c is the number of variables that occur in the second analysis. We then used the per-model average Sorensen's index value as an independent variable to predict the deviation score in a general linear model, and included no random effect since each analysis is included only once, in R (R Core Team, 2024):

$$\text{DeviationScore}_j \sim \beta_{\text{Sorensen}_j} \tag{2.9}$$

Additional Explanation

When we planned this analysis, we anticipated that analysts would identify a single primary effect from each model, so that each model would appear in the analysis only once. Our expecation was incorrect because some analysts identified >1 effect per analysis, but we still chose to specify our model as registered and not use a random effect. This is because most models produced only one effect and so we expected that specifying a random effect to account for the few cases where >1 effect was included for a given model would prevent model convergence.

Note that this analysis contrasts with the analyses in which we used reviewer ratings as predictors because in the analyses with reviewer ratings, each effect appeared in the analysis approximately four times due to multiple reviews of each analysis, and so it was much more important to account for that variance through a random effect.

Next, we assessed the relationship between the inclusion of random effects in the analysis and the deviation from the mean effect size. We anticipated that most analysts would use random effects in a mixed model framework, but if we were wrong, we wanted to evaluate the differences in outcomes when using random effects versus not using random effects. Thus if there were at least 5 analyses that did and 5 analyses that did not include random effects, we would add a binary predictor variable “random effects included (yes/no)” to our set of univariate analyses and would add this predictor variable to our multivariate model described below. This standard was only met for the *Eucalyptus* analyses, and so we only examined inclusion of random effects as a predictor variable in meta-analysis of this set to analyses.

Finally, we conducted a multivariate analysis with the five predictors described above (peer ratings 0-100 and peer ratings of publishability 1-4; both original and revised and Sorensen’s index, plus a sixth for *Eucalyptus*, presence /absence of random effects) with random effects of analysis identity and reviewer identity in the *lme4* package in R (Bates et al., 2015; R Core Team, 2024). We had stated here in the text that we would use only

the revised (final) peer ratings in this analysis, so the absence of the initial ratings is not a deviation from our plan:

$$\begin{aligned}
\text{DeviationScore}_j &= \text{BoxCox}(|\text{DeviationFromMean}_j|) \\
\text{DeviationScore}_{ij} &\sim \text{RatingContinuous}_{ij} + \\
&\quad \text{RatingCategorical}_{ij} + \\
&\quad \beta_{\text{Sorensen}\,j} + \\
&\quad \text{AnalysisID}_j + \\
&\quad \text{ReviewerID}_i \\
\text{ReviewerID}_i &\sim \mathcal{N}(0, \sigma_i^2) \\
\text{AnalysisID}_j &\sim \mathcal{N}(0, \sigma_j^2)
\end{aligned} \tag{2.10}$$

We conducted all the analyses described above eight times; for each of the four responses (Z_r, y_{25th}, y_{50th}, y_{75th}) one time for each of the two datasets.

We have publicly archived all relevant data, code, and materials on the Open Science Framework (osf.io/mn5aj). Archived data includes the original datasets distributed to all analysts, any edited versions of the data analyzed by individual groups, and the data we analyzed with our meta-analyses, which include the effect sizes derived from separate analyses, the statistics describing variation in model structure among analyst groups, and the anonymised answers to our surveys of analysts and peer reviewers. Similarly, we have archived both the analysis code used for each individual analysis (where available) and the code from our meta-analyses. We have also archived copies of our survey instruments from analysts and peer reviewers.

Our rules for excluding data from our study were as follows. We excluded from our synthesis any individual analysis submitted after we had completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We also excluded any individual analysis that did not produce an outcome that could be interpreted as an answer to our primary question (as posed above) for the respective dataset. For instance, this means that in the case of the data on blue tit chick growth, we excluded any analysis that did not include something that can be interpreted as growth or size as a dependent (response) variable, and in the case of the *Eucalyptus* establishment data, we

excluded any analysis that did not include a measure of grass cover among the independent (predictor) variables. Also, as described above, any analysis that could not produce an effect that could be converted to a signed Z_r was excluded from analyses of Z_r.

Preregistration Deviation

Some analysts had difficulty implementing our instructions to derive the out-of-sample predictions, and in some cases (especially for the *Eucalyptus* data), they submitted predictions with implausibly extreme values. We believed these values were incorrect and thus made the conservative decision to exclude out-of-sample predictions where the estimates were > 3 standard deviations from the mean value from the full dataset provided to teams for analysis.

Additional Explanation: we conducted several unregistered analyses

1. Evaluating model fit.

We evaluated all fitted models using the `performance()` function from the *performance* package (Lüdecke, Ben-Shachar, et al., 2021) and the `glance()` function from the *broom.mixed* package (Bolker et al., 2024). For all models, we calculated the square root of the residual variance (Sigma) and the root mean squared error (RMSE). For GLMMs `performance()` calculates the marginal and conditional R^2 values as well as the contribution of random effects (ICC), based on Nakagawa et al. (2017). The conditional R^2 accounts for both the fixed and random effects, while the marginal R^2 considers only the variance of the fixed effects. The contribution of random effects is obtained by subtracting the marginal R^2 from the conditional R^2.

2. Exploring outliers and analysis quality.

After seeing the forest plots of Z_r values and noticing the existence of a small number of extreme outliers, especially from the *Eucalyptus* analyses, we wanted to understand the degree to which our heterogeneity estimates were influenced by these outliers. To explore this question, we removed the highest two and lowest two values of Z_r in each dataset and re-calculated our heterogeneity estimates.

To help understand the possible role of the quality of analyses in driving the heterogeneity we observed among estimates of Z_r, we created forest plots and recalculated our heterogeneity estimates after removing all effects from analysis teams that had received at least one rating of "deeply flawed and unpublishable" and then again after

removing all effects from analysis teams with at least one rating of either "deeply flawed and unpublishable" or "publishable with major revisions". We also used self-identified levels of statistical expertise to examine heterogeneity when we retained analyses only from analysis teams that contained at least one member who rated themselves as "highly proficient" or "expert" (rather than "novice" or "moderately proficient") in conducting statistical analyses in their research area in our intake survey.

Additionally, to assess potential impacts of highly collinear predictor variables on estimates of Z_r in blue tit analyses, we created forest plots (Figure A.8) and recalculated our heterogeneity estimates after we removed analyses that contained the brood count after manipulation and the highly correlated (correlation of 0.89, Appendix A.4) brood count at day 14. This removal included the one effect based on a model that contained both these variables and a third highly correlated variable, the estimate of number of chicks fledged (the only model that included the estimate of number of chicks fledged). We did not conduct a similar analysis for the *Eucalyptus* dataset because there were no variables highly collinear with the primary predictors (grass cover variables) in that dataset (Appendix A.4).

3. Exploring possible impacts of lower quality estimates of degrees of freedom. Our meta-analyses of variation in Z_r required variance estimates derived from estimates of the degrees of freedom in original analyses from which Z_r estimates were derived. While processing the estimates of degrees of freedom submitted by analysts, we identified a subset of these estimates in which we had lower confidence because two or more effects from the same analysis were submitted with identical degrees of freedom. We therefore conducted a second set of (more conservative) meta-analyses that excluded these Z_r estimates with identical estimates of degrees of freedom and we present these analyses in the supplement.

Additional Explanation: best practices in many-analysts research

After we initiated our project, a paper was published outlining best practices in many-analysts studies (Aczel et al., 2021). Although we did not have access to this document when we implemented our project, our study complies with these practices nearly completely. The one exception is that although we requested analysis code from analysts, we did not require submission of code.

2.2.6 Step 6: Facilitated Discussion and Collaborative Write-Up of Manuscript

We planned for analysts and initiating authors to discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

> **Preregistration Deviation**
> As described above, due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We built an R package, *ManyEcoEvo* to conduct the analyses described in this study (Gould et al., 2023), which can be downloaded from github.com/egouldo/ManyEcoEvo to reproduce our analyses or replicate the analyses described here using alternate datasets. Data cleaning and preparation of analysis-data, as well as the analysis, is conducted in R (R Core Team, 2024) reproducibly using the *targets* package (Landau, 2021). This data and analysis pipeline is stored in the *ManyEcoEvo* package repository and its outputs are made available to users of the package when the library is loaded.

The full manuscript, including further analysis and presentation of results is written in Quarto (Allaire & Dervieux, 2024). The source code to reproduce the manuscript is hosted at github.com/egouldo/ManyAnalysts (Gould et al., 2024), and the rendered version of the source code may be viewed at egouldo.github.io/ManyAnalysts. All R packages and their versions used in the production of the manuscript are listed in Appendix A.5.

2.3 Results

2.3.1 Summary Statistics

In total, 173 analyst teams, comprising 246 analysts, contributed 182 usable analyses (compatible with our meta-analyses and provided with all information needed for inclusion) of the two datasets examined in this study which yielded 215 effects. Analysts produced 134 distinct effects that met our criteria for inclusion in at least one of our meta-analyses for the blue tit dataset. Analysts produced 81 distinct effects meeting our criteria for inclusion for the *Eucalyptus* dataset. Excluded analyses and effects either did not answer our specified

biological questions, were submitted with insufficient information for inclusion in our meta-analyses, or were incompatible with production of our effect size(s). We expected cases of this final scenario (incompatible analyses), for instance we cannot extract a Z_r from random forest models, which is why we analyzed two distinct types of effects, Z_r and out-of-sample. Some effects only provided sufficient information for a subset of analyses and were only included in that subset. For both datasets, most submitted analyses incorporated mixed effects. Submitted analyses of the blue tit dataset typically specified normal error and analyses of the *Eucalyptus* dataset typically specified a non-normal error distribution (Table A.1).

For both datasets, the composition of models varied substantially in regards to the number of fixed and random effects, interaction terms, and the number of data points used, and these patterns differed somewhat between the blue tit and *Eucalyptus* analyses (See Table A.2). Focussing on the models included in the Z_r analyses (because this is the larger sample), blue tit models included a similar number of fixed effects on average (mean 5.2 $\pm$ 2.92 SD, range: 1 to 19) as *Eucalyptus* models (mean 5.01 $\pm$ 3.83 SD, range: 1 to 13), but the standard deviation in number of fixed effects was somewhat larger in the *Eucalyptus* models. The average number of interaction terms was much larger for the blue tit models (mean 0.44 $\pm$ 1.11 SD, range: 0 to 10) than for the *Eucalyptus* models (mean 0.16 $\pm$ 0.65 SD, range: 0 to 5), but still under 0.5 for both, indicating that most models did not contain interaction terms. Blue tit models also contained more random effects (mean 3.53 $\pm$ 2.08 SD, range: 0 to 10) than *Eucalyptus* models (mean 1.41 $\pm$ 1.09 SD, range: 0 to 4). The maximum possible sample size in the blue tit dataset (3720 nestlings) was an order of magnitude larger than the maximum possible in the *Eucalyptus* dataset (351 plots), and the means and standard deviations of the sample size used to derive the effects eligible for our study were also an order of magnitude greater for the blue tit dataset (mean 2611.09 $\pm$ 937.48 SD, range: 76 to 76) relative to the *Eucalyptus* models (mean 298.43 $\pm$ 106.25 SD, range: 18 to 351). However, the standard deviation in sample size from the *Eucalyptus* models was heavily influenced by a few cases of dramatic sub-setting (described below). Approximately three quarters of *Eucalyptus* models used sample sizes within 3% of the maximum. In contrast, fewer than 20% of blue tit models relied on sample sizes within 3% of the maximum, and approximately 50% of blue tit models relied on sample sizes 29% or more below the maximum.

Analysts provided qualitative descriptions of the conclusions of their analyses. Each analysis team provided one conclusion per dataset. These conclusions could take into account the results of any formal analyses completed by the team as well as exploratory and visual analyses of the data. Here we summarize all qualitative responses, regardless of whether we had sufficient information to use the corresponding model results in our quantitative analyses below. We classified these conclusions into the categories summarised below (Table 2.1):

- *Mixed*: some evidence supporting a positive effect, some evidence supporting a negative effect
- *Conclusive negative*: negative relationship described without caveat
- *Qualified negative*: negative relationship but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive none*: analysts interpret the results as conclusive of no effect
- *None qualified*: analysts describe finding no evidence of a relationship but they describe the potential for an undetected effect
- *Qualified positive*: positive relationship described but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive positive*: positive relationship described without caveat

For the blue tit dataset, most analysts concluded that there was negative relationship between measures of sibling competition and nestling growth, though half the teams expressed qualifications or described effects as mixed or absent. No analysts concluded that there was a positive relationship even though some individual effect sizes were positive, apparently because all analysts who produced effects indicating positive relationships also produced effects indicating negative relationships and therefore described their results as qualified, mixed, or absent. For the *Eucalyptus* dataset, there was a broader spread of conclusions with at least one analyst team providing conclusions consistent with each conclusion category. The most common conclusion for the *Eucalyptus* dataset was that there was no relationship between grass cover and *Eucalyptus* recruitment (either conclusive or qualified description of no relationship), but more than half the teams concluded that there were effects; negative, positive, or mixed.

Table 2.1: *Tallies of analysts' qualitative answers to the research questions addressed by their analyses.*

Dataset	Mixed	Negative Conclusive	Negative Qualified	None Conclusive	None Qualified	Positive Conclusive	Positive Qualified
blue tit	5	37	27	4	1	0	0
Eucalyptus	8	6	12	19	12	2	4

2.3.2 Distribution of Effects

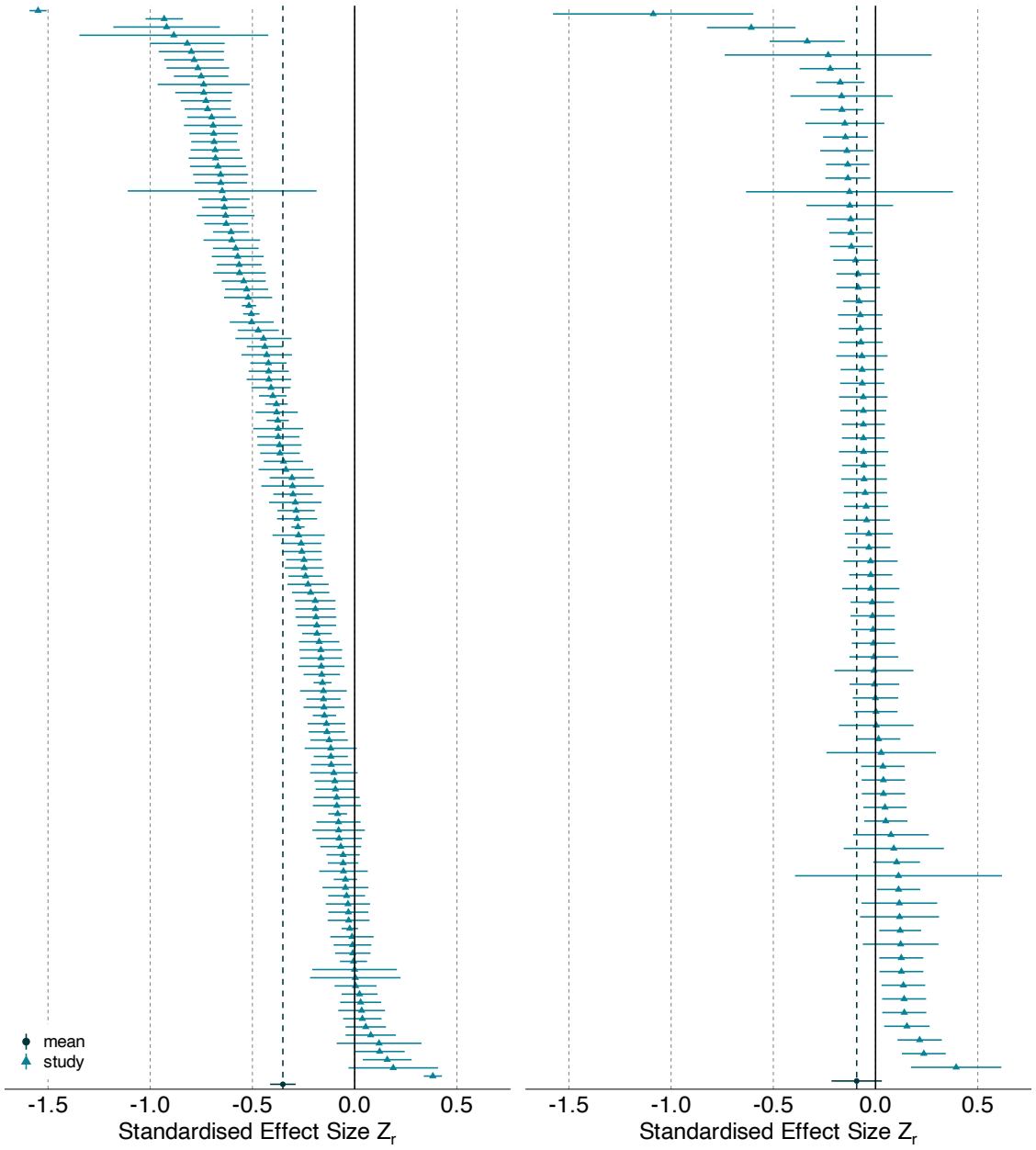
Standardised Effect Sizes (Z_r)

Although the majority (118 of 131) of the usable Z_r effects from the blue tit dataset found nestling growth decreased with sibling competition, and the meta-analytic mean $\bar{Z}_r$ (Fisher's transformation of the correlation coefficient) was convincingly negative (-0.35 $\pm$ 0.06 95%CI), there was substantial variability in the strength and the direction of this effect. Z_r ranged from -1.55 to 0.38, and approximately continuously from -0.93 to 0.19 (Figure 2.2a and Table 2.4), and of the 118 effects with negative slopes, 93 had confidence intervals excluding 0. Of the 13 with positive slopes indicating increased nestling growth in the presence of more siblings, 2 had confidence intervals excluding zero (Figure 2.2a).

Meta-analysis of the *Eucalyptus* dataset also showed substantial variability in the strength of effects as measured by Z_r, and unlike with the blue tits, a notable lack of consistency in the direction of effects (Figure 2.2b, Table 2.4). Z_r ranged from -4.47 (Figure A.2), indicating a strong tendency for reduced *Eucalyptus* seedling success as grass cover increased, to 0.39, indicating the opposite. Although the range of reported effects skewed strongly negative, this was due to a small number of substantial outliers. Most values of Z_r were relatively small with values $> |0.2|$ and the meta-analytic mean effect size was close to zero (-0.09 $\pm$ 0.12 95% CI). Of the 79 effects, fifty-three had confidence intervals overlapping zero, approximately a quarter (fifteen) crossed the traditional threshold of statistical significance indicating a negative relationship between grass cover and seedling success, and eleven crossed the significance threshold indicating a positive relationship between grass cover and seedling success (Figure 2.2b).

Out-of-sample Predictions y_i

As with the effect size Z_r, we observed substantial variability in the size of out-of-sample predictions derived from the analysts' models. Blue tit predictions (Figure 2.3), which were z-score-standardised to accommodate the use of different response variables, always ranged



(a) *Blue tit analyses: Points where Z_r are less than 0 indicate analyses that found a negative relationship between sibling number and nestling growth.*

(b) Eucalyptus *analyses: Points where Z_r are less than 0 indicate a negative relationship between grass cover and* Eucalyptus *seedling success.*

Figure 2.2: *Forest plots of meta-analytic estimated standardised effect sizes (Z_r, blue triangles) and their 95% confidence intervals for each effect size included in the meta-analysis model. The meta-analytic mean effect size is denoted by a black circle and a dashed vertical line, with error bars also representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0, indicating no relationship between the test variable and the response variable. Note that the* Eucalyptus *plot omits one extreme outlier with the value of -4.47 (Figure A.2) in order to standardize the x-axes on these two panels.*

far in excess of one standard deviation. In the y_{25} scenario, model predictions ranged from -1.84 to 0.42 (a range of 2.68 standard deviations), in the y_{50} they ranged from -0.52 to 1.08 (a range of 1.63 standard deviations), and in the y_{75} scenario they ranged from -0.03 to 1.59 (a range of 1.9 standard deviations). As should be expected given the existence of both negative and positive Z_r values, all three out-of-sample scenarios produced both negative and positive predictions, although as with the Z_r values, there is a clear trend for scenarios with more siblings to be associated with smaller nestlings. This is supported by the meta-analytic means of these three sets of predictions which were -0.66 (95%CI -0.82–0.5) for the y_{25}, 0.34 (95%CI 0.2-0.48) for the y_{50}, and 0.67 (95%CI 0.57-0.77) for the y_{75}.

Eucalyptus out-of-sample predictions also varied substantially (Figure 2.4), but because they were not z-score-standardised and are instead on the original count scale, the types of interpretations we can make differ. The predicted *Eucalyptus* seedling counts per 15 x 15 m plot for the y_{25} scenario ranged from 0.04 to 26.68, for the y_{50} scenario ranged from 0.04 to 44.31, and for the y_{75} scenario they ranged from 0.03 to 60.95. The meta-analytic mean predictions for these three scenarios were similar; 1.28 (95%CI 0.59-2.29) for the y_{25}, 2.92 (95%CI 0.98-3.89) for the y_{50}, and 2.92 (95%CI 1.59-4.9) for the y_{75} scenarios respectively.

2.3.3 Quantifying Heterogeneity

Effect Sizes (Z_r)

We quantified both absolute (τ^2) and relative (I^2) heterogeneity resulting from analytical variation. Both measures suggest that substantial variability among effect sizes was attributable to the analytical decisions of analysts.

The total absolute level of variance beyond what would typically be expected due to sampling error, τ^2 (Table 2.2), among all usable blue tit effects was 0.08 and for *Eucalyptus* effects was 0.27. This is similar to or exceeding the median value (0.105) of τ^2 found across 31 recent meta-analyses (calculated from the data in Yang et al., 2023). The similarity of our observed values to values from meta-analyses of different studies based on different data suggest the potential for a large portion of heterogeneity to arise from analytical decisions. For further discussion of interpretation of τ^2 in our study, please consult discussion of *post hoc* analyses below.

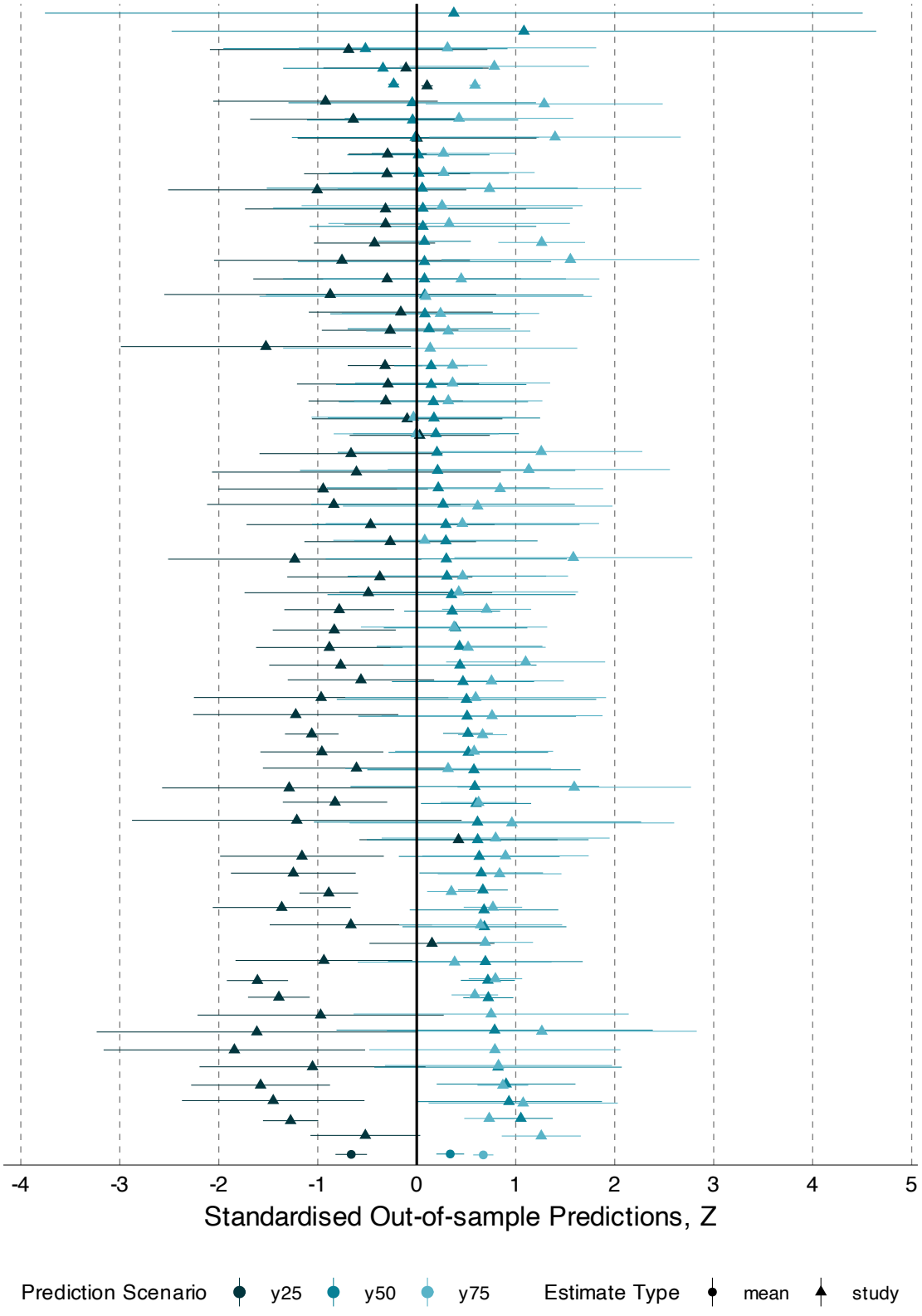


Figure 2.3: *Forest plot of meta-analytic estimated standardised (z-score) blue tit out-of-sample predictions, y_i. Triangles represent individual estimates. Circles represent the meta-analytic mean for each prediction scenario. Dark-blue points correspond to y_{25} scenario, medium-blue points correspond to the y_{50} scenario, while light blue points correspond to the y_{75} scenario. Error bars are 95% confidence intervals.*

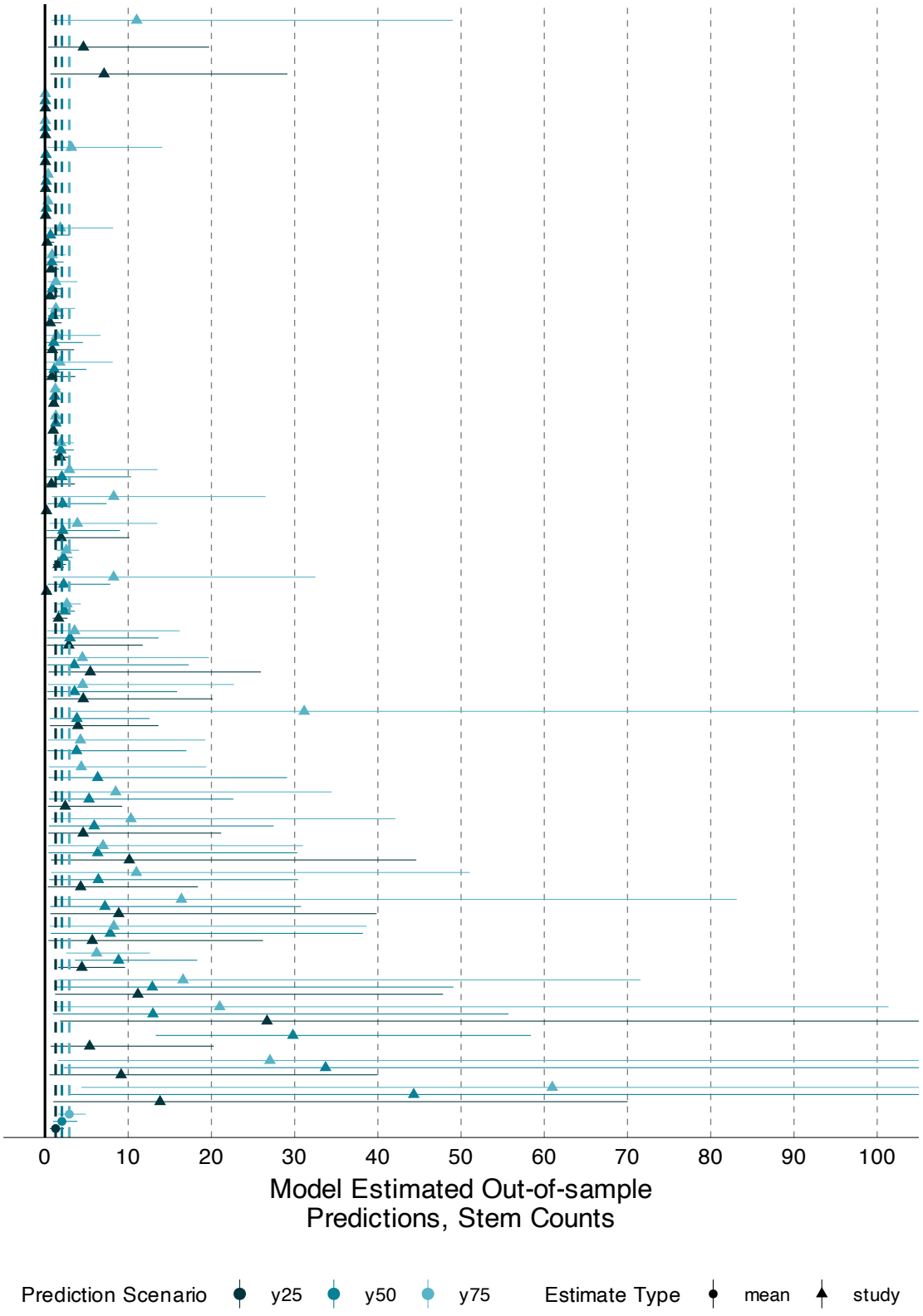


Figure 2.4: *Forest plot of meta-analytic estimated* Eucalyptus *out-of-sample predictions, y_i, on the response-scale (stems counts). Triangles represent individual estimates. Circles represent the meta-analytic mean for each prediction scenario. Dark-blue points correspond to y_{25} scenario, medium-blue points correspond to the y_{50} scenario, while light blue points correspond to the y_{75} scenario. Error bars are 95% confidence intervals. Outliers (observations more than 3SD above the mean) have been removed prior to model fitting and do not appear on this figure. x-axis is truncated to 100, and thus some error bars are incomplete. See Figure A.9 for full figure.*

Table 2.2: *Heterogeneity in the estimated effects Z_r for meta-analyses of: the full dataset, as well as from post hoc analyses wherein analyses with outliers are removed, analyses with effects from analysis teams with at least one "unpublishable" rating are excluded, analyses receiving at least one "major revisions" rating or worse excluded, analyses from teams with at least one analyst self-rated as "highly proficient" or "expert" in statistical analysis are included, and (blue tit only) analyses that did not included the pair of highly collinear predictors together. τ^2_{Team} is the absolute heterogeneity for the random effect Team. $\tau^2_{effectID}$ is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{Team,\ effectID}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences.*

Dataset	N_{Obs}	τ^2_{Total}	τ^2_{Team}	$\tau^2_{effectID}$	I^2_{Total}	I^2_{Team}	$I^2_{Team,\ effectID}$
All analyses							
blue tit	131	0.08	0.03	0.05	97.61%	36.71%	60.90%
Eucalyptus	79	0.27	0.02	0.25	98.59%	6.89%	91.70%
Blue tit analyses containing highly collinear predictors removed							
blue tit	117	0.07	0.04	0.03	96.92%	58.18%	38.75%
All analyses, Outliers removed							
blue tit	127	0.07	0.04	0.02	96.84%	64.63%	32.21%
Eucalyptus	75	0.01	0.00	0.01	66.19%	19.25%	46.94%
Analyses receiving at least one 'Unpublishable' rating removed							
blue tit	109	0.08	0.03	0.05	97.52%	35.68%	61.84%
Eucalyptus	55	0.01	0.01	0.01	79.74%	28.31%	51.43%
Analyses receiving at least one 'Unpublishable' and or 'Major Revisions' rating removed							
blue tit	32	0.14	0.01	0.13	98.72%	5.17%	93.55%
Eucalyptus	13	0.03	0.03	0.00	88.91%	88.91%	0.00%
Analyses from teams with highly proficient or expert data analysts							
blue tit	89	0.09	0.03	0.06	97.91%	31.43%	66.49%
Eucalyptus	34	0.58	0.02	0.56	99.41%	3.47%	95.94%

In our analyses, I^2 is a plausible index of how much more variability among effect sizes we have observed, as a proportion, than we would have observed if sampling error were driving variability. We discuss our interpretation of I^2 further in the methods, but in short, it is a useful metric for comparison to values from published meta-analyses and provides a plausible value for how much heterogeneity could arise in a normal meta-analysis with similar sample sizes due to analytical variability alone. In our study, total I^2 for the blue tit Z_r estimates was extremely large, at 97.61%, as was the *Eucalyptus* estimate (98.59% Table 2.2).

Although the overall I^2 values were similar for both *Eucalyptus* and blue tit analyses, the relative composition of that heterogeneity differed. For both datasets, the majority of heterogeneity in Z_r was driven by differences among effects as opposed to differences among teams, though this was more prominent for the *Eucalyptus* dataset, where nearly all of the total heterogeneity was driven by differences among effects (91.7%) as opposed to differences among teams (6.89%) (Table 2.2).

Out-of-sample Predictions (y_i)

We observed substantial heterogeneity among out-of-sample estimates, but the pattern differed somewhat from the Z_r values (Table 2.3). Among the blue tit predictions, I^2 ranged from medium-high for the y_{25} scenario (68.54) to low (27.9) for the y_{75} scenario. Among the *Eucalyptus* predictions, I^2 values were uniformly high (>82%). For both datasets, most of the existing heterogeneity among predicted values was attributable to among-team differences, with the exception of the y_{50} analysis of the *Eucalyptus* dataset. We are limited in our interpretation of τ^2 for these estimates because, unlike for the Z_r estimates, we have no benchmark for comparison with other meta-analyses.

2.3.4 Post-hoc Analysis: Exploring outlier characteristics and the effect of outlier removal on heterogeneity

Effect Sizes (Z_r)

The outlier *Eucalyptus* Z_r values were striking and merited special examination. The three negative outliers had very low sample sizes were based on either small subsets of the dataset or, in one case, extreme aggregation of data. The outliers associated with small subsets had sample sizes (n = 117, 90, 18) that were less than half of the total possible sample size of

Table 2.3: *Heterogeneity among the out-of-sample predictions y_i for both blue tit and* Eucalyptus *datasets. τ^2_{Team} is the absolute heterogeneity for the random effect Team. $\tau^2_{effectID}$ is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{Team,effectID}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences.*

Prediction Scenario		N_{Obs}	τ_{Total}	τ^2_{T}
blue tit				
	y_{25}	63	0.23	
	y_{50}	60	0.23	0
	y_{75}	63	0.23	0
Eucalyptus				
	y_{25}	38	5.75	1
	y_{50}	38	5.75	1
	y_{75}	38	5.75	1

351. The case of extreme aggregation involved averaging all values within each of the 351 sites in the dataset.

Surprisingly, both the largest and smallest effect sizes in the blue tit analyses (Figure 2.2a) come from the same analyst (anonymous ID: 'Adelong'), with identical models in terms of the explanatory variable structure, but with different response variables. However, the radical change in effect was primarily due to collinearity with covariates. The primary predictor variable (brood count after manipulation) was accompanied by several collinear variables, including the highly collinear (correlation of 0.89 Appendix A.4) covariate (brood count at day 14) in both analyses. In the analysis of nestling weight, brood count after manipulation showed a strong positive partial correlation with weight after controlling for brood count at day 14 and treatment category (increased, decreased, unmanipulated). In that same analysis, the most collinear covariate (the day 14 count) had a negative partial correlation with weight. In the analysis with tarsus length as the response variable, these partial correlations were almost identical in absolute magnitude, but reversed in sign and so brood count after manipulation was now the collinear predictor with the negative relationship. The two models were therefore very similar, but the two collinear predictors simply switched roles, presumably because a subtle difference in the distribution of weight and tarsus length data.

When we dropped the *Eucalyptus* outliers, I^2 decreased from high (98.59 %), using Higgins' (Higgins et al., 2003) suggested benchmark, to between moderate and high (66.19 %, Table 2.2). However, more notably, τ^2 dropped from 0.27 to 0.01, indicating that, once outliers were excluded, the observed variation in effects was similar to what we would expect if sampling error were driving the differences among effects (since τ^2 is the variance in addition to that driven by sampling error). The interpretation of this value of τ^2 in the context of our many-analyst study is somewhat different than a typical meta-analysis, however, since in our study (especially for *Eucalyptus*, where most analyses used almost exactly the same data points), there is almost no role for sampling error in driving the observed differences among the estimates. Thus, rather than concluding that the variability we observed among estimates (after removing outliers) was due only to sampling error (because τ^2 became small: 10% of the median from Yang et al., 2023), we instead conclude that the observed variability, which must be due to the divergent choices of analysts rather than sampling error, is approximately of the same magnitude as what we would have expected if, instead, sampling error, and not analytical heterogeneity, were at work. Presumably, if sampling error had actually also been at work, it would have acted as an additional source of variability and would have led total variability among estimates to be higher. With total variability higher and thus greater than expected due to sampling error alone, τ^2 would have been noticeably larger. Conversely, dropping outliers from the set of blue tit effects did not meaningfully reduce I^2 , and only modestly reduced τ^2 (Table 2.2). Thus, effects at the extremes of the distribution were much stronger contributors to total heterogeneity for effects from analyses of the *Eucalyptus* than for the blue tit dataset.

Out-of-sample Predictions (y_i)

We did not conduct these *post hoc* analyses on the out-of-sample predictions as the number of eligible effects was smaller and the pattern of outliers differed.

2.3.5 Post-hoc Analysis: Exploring the effect of removing analyses with poor peer ratings on heterogeneity

Effect Sizes (Z_r)

Removing poorly rated analyses had limited impact on the meta-analytic means (Figure A.6). For the *Eucalyptus* dataset, the meta-analytic mean shifted from -0.09 to -0.02 when effects from analyses rated as unpublishable were removed, and to -0.04 when effects from

Table 2.4: *Estimated mean value of the standardised correlation coefficient, Z_r, along with its standard error and 95% confidence intervals. We re-computed the meta-analysis for different post-hoc subsets of the data: All eligible effects, removal of effects from blue tit analyses that contained a pair of highly collinear predictor variables, removal of effects from analysis teams that received at least one peer rating of "deeply flawed and unpublishable", removal of any effects from analysis teams that received at least one peer rating of either "deeply flawed and unpublishable" or "publishable with major revisions", inclusion of only effects from analysis teams that included at least one member who rated themselves as "highly proficient" or "expert" at conducting statistical analyses in their research area.*

Dataset	$\hat{\mu}$	SE[$\hat{\mu}$]	95% CI	statistic	p
All analyses					
blue tit	-0.35	0.03	[-0.41,-0.29]	-11.02	<0.001
Eucalyptus	-0.09	0.06	[-0.22,0.03]	-1.47	0.14
Blue tit analyses containing highly collinear predictors removed					
blue tit	-0.36	0.03	[-0.42,-0.29]	-10.97	<0.001
All analyses, outliers removed					
blue tit	-0.36	0.03	[-0.42,-0.30]	-11.48	<0.001
Eucalyptus	-0.03	0.01	[-0.06,0.00]	-2.23	0.026
Analyses receiving at least one 'Unpublishable' rating removed					
blue tit	-0.36	0.03	[-0.43,-0.30]	-10.82	<0.001
Eucalyptus	-0.02	0.02	[-0.07,0.02]	-1.15	0.3
Analyses receiving at least one 'Unpublishable' and or 'Major Revisions' rating removed					
blue tit	-0.37	0.07	[-0.51,-0.23]	-5.34	<0.001
Eucalyptus	-0.04	0.05	[-0.15,0.07]	-0.77	0.4
Analyses from teams with highly proficient or expert data analysts					
blue tit	-0.36	0.04	[-0.44,-0.28]	-8.93	<0.001
Eucalyptus	-0.17	0.13	[-0.43,0.10]	-1.24	0.2

analyses rated, at least once, as unpublishable or requiring major revisions were removed. Further, the confidence intervals for all of these means overlapped each of the other means (Table 2.4). We saw similar patterns for the blue tit dataset, with only small shifts in the meta-analytic mean, and confidence intervals of all three means overlapping each other mean (Table 2.4). Refitting the meta-analysis with a fixed effect for categorical ratings also showed no indication of differences in group meta-analytic means due to peer ratings (Figure A.3).

For the blue tit dataset, removing poorly-rated analyses led to only negligible changes in I^2_{Total} and relatively minor impacts on τ^2. However, for the *Eucalyptus* dataset, removing poorly-rated analyses led to notable reductions in I^2_{Total} and substantial reductions in τ^2. When including all analyses, the *Eucalyptus* I^2_{Total} was 98.59% and τ^2 was 0.27, but eliminating analyses with ratings of "unpublishable" reduced I^2_{Total} to 79.74% and τ^2 to 0.01, and removing also those analyses "needing major revisions" left I^2_{Total} at 88.91% and τ^2 at 0.03 (Table 2.2). Additionally, the allocations of I^2 to the team versus individual effect were altered for both blue tit and *Eucalyptus* meta-analyses by removing poorly rated analyses,

but in different ways. For blue tit meta-analysis, between a third and two-thirds of the total I^2 was attributable to among-team variance in most analyses until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost all remaining heterogeneity was attributable to among-effect differences. In contrast, for *Eucalyptus* meta-analysis, the among-team component of I^2 was less than third until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost 90% of heterogeneity was attributable to differences among teams.

Out-of-sample Predictions y_i

We did not conduct these *post hoc* analyses on the out-of-sample predictions as the number of eligible effects was smaller and our ability to interpret heterogeneity values for these analyses was limited.

2.3.6 Post-hoc Analysis: Exploring the effect of including only analyses conducted by analysis teams with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area

Effect Sizes (Z_r)

Including only analyses conducted by teams that contained at least one member who rated themselves as "highly proficient" or "expert" in conducting the relevant statistical methods had negligible impacts on the meta-analytic means (Table 2.4), the distribution of Z_r effects (Figure A.7), or heterogeneity estimates (Table 2.2), which remained extremely high.

Out-of-sample Predictions y_i

We did not conduct these *post hoc* analyses on the out-of-sample predictions as the number of eligible effects was smaller.

2.3.7 Post-hoc Analysis: Exploring the effect of excluding estimates of Z_r in which we had reduced confidence

As described in our addendum to the methods, we identified a subset of estimates of Z_r in which we had less confidence because of features of the submitted degrees of freedom. Excluding these effects in which we had lower confidence had minimal impact on the

Table 2.5: *Summary metrics for registered models seeking to explain deviation (Box-Cox transformed absolute deviation scores) from the mean Z_r as a function of Sorensen's index, categorical peer ratings, and continuous peer ratings for blue tit and* Eucalyptus *analyses, and as a function of the presence or absence of random effects (in the analyst's models) for* Eucalyptus *analyses. We report coefficient of determination, R^2, for our models including only fixed effects as predictors of deviation, and we report $R^2_{Conditional}$, $R^2_{Marginal}$ and the intra-class correlation (ICC) from our models that included both fixed and random effects. For all our models, we calculated the residual standard deviation σ and root mean squared error (RMSE).*

Dataset	$N_{Obs.}$	R^2	$R^2_{Conditional}$	$R^2_{Marginal}$	ICC	σ	RMSE
Deviation explained by categorical ratings							
blue tit	473		0.09	7.47×10^{-3}	0.08	0.5	0.48
Eucalyptus	346		0.13	0.01	0.12	1.06	1.02
Deviation explained by continuous ratings							
blue tit	473		0.09	3.44×10^{-3}	0.09	0.5	0.48
Eucalyptus	346		0.12	7.44×10^{-3}	0.11	1.06	1.03
Deviation explained by Sorensen's index							
blue tit	131	6.32×10^{-3}				0.51	0.51
Eucalyptus	79	1.84×10^{-4}				1.12	1.1
Deviation explained by inclusion of random effects							
Eucalyptus	79	8.75×10^{-8}				1.12	1.1

meta-analytic mean and the estimates of total I^2 and τ^2 for both blue tit and *Eucalyptus* meta-analyses, regardless of whether outliers were also excluded (Table A.4).

2.3.8 Post-hoc Analysis: Exploring the effect of excluding effects from blue tit models that contained two highly collinear predictors

Effect Sizes (Z_r)

Excluding effects from blue tit models that contained the two highly collinear predictors (brood count after manipulation and brood count at day 14) had negligible impacts on the meta-analytic means (Table 2.4), the distribution of Z_r effects (Figure A.8), or heterogeneity estimates (Table 2.2), which remained high.

Out-of-sample Predictions y_i

Inclusion of collinear predictors does not harm model prediction, and so we did not conduct these *post-hoc* analyses.

2.3.9 Explaining Variation in Deviation Scores

None of the pre-registered predictors explained substantial variation in deviation among submitted statistical effects from the meta-analytic mean (Table 2.5, Table 2.6).

Table 2.6: *Parameter estimates from models of Box-Cox transformed deviation scores from the mean Z_r as a function of continuous and categorical peer ratings, Sorensen scores, and the inclusion of random effects. Standard Errors (SE), 95% confidence intervals (95% CI) are reported for all estimates, while t values, degrees of freedom and p-values are presented for fixed-effects. Note that positive parameter estimates mean that as the predictor variable increases, so does the absolute value of the deviation from the meta-analytic mean.*

Parameter	Random Effect	Coefficient	SE	95% CI	t	df	*p*
Deviation explained by inclusion of random effects - *Eucalyptus*							
(Intercept)		-2.53	0.27	[-3.06, -1.99]	-9.31	77	<0.001
Mixed model		0.00	0.31	[-0.60, 0.60]	0.00	77	>0.9
Deviation explained by Sorensen's index - blue tit							
(Intercept)		-1.53	0.28	[-2.08, -0.98]	-5.42	129	<0.001
Mean Sorensen's index		0.42	0.47	[-0.49, 1.34]	0.91	129	0.4
Deviation explained by Sorensen's index - *Eucalyptus*							
(Intercept)		-2.65	1.05	[-4.70, -0.60]	-2.53	77	0.011
Mean Sorensen's index		0.18	1.51	[-2.78, 3.14]	0.12	77	>0.9
Deviation explained by continuous ratings - blue tit							
(Intercept)		-1.16	0.11	[-1.37, -0.94]	-10.60	469	<0.001
RateAnalysis		-0.002	0	[-0.004, 0]	-1.22	469	0.2
SD (Intercept)	Reviewer ID	0.16	0.03	[0.10, 0.24]			
SD (Observations)	Residual	0.5	0.02	[0.46, 0.53]			
Deviation explained by continuous ratings - *Eucalyptus*							
(Intercept)		-2.23	0.23	[-2.69, -1.78]	-9.65	342	<0.001
RateAnalysis		-0.004	0	[-0.011, 0]	-1.44	342	0.15
SD (Intercept)	Reviewer ID	0.37	0.09	[0.24, 0.60]			
SD (Observations)	Residual	1.06	0.04	[0.98, 1.15]			
Deviation explained by categorical ratings - blue tit							
(Intercept)		-1.11	0.11	[-1.33, -0.89]	-9.91	467	<0.001
Publishable with major revision		-0.19	0.12	[-0.42, 0.04]	-1.62	467	0.10
Publishable with minor revision		-0.19	0.12	[-0.42, 0.04]	-1.65	467	0.10
Publishable as is		-0.13	0.13	[-0.39, 0.12]	-1.02	467	0.3
SD (Intercept)	Reviewer ID	0.15	0.04	[0.10, 0.24]			
SD (Observations)	Residual	0.5	0.02	[0.46, 0.53]			
Deviation explained by categorical ratings - *Eucalyptus*							
(Intercept)		-2.66	0.27	[-3.18, -2.13]	-9.97	340	<0.001
Publishable with major revision		0.29	0.29	[-0.27, 0.85]	1.02	340	0.3
Publishable with minor revision		0.01	0.28	[-0.54, 0.56]	0.04	340	>0.9
Publishable as is		0.05	0.31	[-0.55, 0.66]	0.17	340	0.9
SD (Intercept)	Reviewer ID	0.39	0.09	[0.25, 0.61]			
SD (Observations)	Residual	1.06	0.04	[0.98, 1.15]			

2.3.10 Deviation Scores as Explained by Reviewer Ratings

Effect Sizes (Z_r)

We obtained reviews from 153 reviewers who reviewed analyses for a mean of 3.27 (range 1 - 11) analysis teams. Analyses of the blue tit dataset received a total of 240 reviews, each was reviewed by a mean of 3.87 (SD 0.71, range 3-5) reviewers. Analyses of the *Eucalyptus* dataset received a total of 178 reviews, each was reviewed by a mean of 4.24 (SD 0.79, range 3-6) reviewers. We tested for inter-rater-reliability to examine how similarly reviewers reviewed each analysis and found approximately no agreement among reviewers. When considering continuous ratings, IRR was 0.01, and for categorical ratings, IRR was -0.14.

Many of the models of deviation as a function of peer ratings faced issues of failure to converge or singularity due to sparse design matrices with our pre-registered random effects (Effect ID and Reviewer ID) (see supplementary material A.6). These issues persisted after increasing the tolerance and changing the optimizer. For both *Eucalyptus* and blue tit datasets, models with continuous ratings as a predictor were singular when both pre-registered random effects were included.

When using only categorical ratings as predictors, models converged only when specifying Reviewer ID as a random effect. That model had a R^2_C of 0.09 and a R^2_M of 0.01. The model using the continuous ratings converged for both random effects (in isolation), but not both. We present results for the model using study ID as a random effect because we expected it would be a more important driver of variation in deviation scores. That model had a R^2_C of 0.09 and a R^2_M of 0.01 for the blue tit dataset and a R^2_C of 0.12 and a R^2_M of 0.01 for the *Eucalyptus* dataset. Neither continuous or categorical reviewer ratings of the analyses meaningfully predicted deviance from the meta-analytic mean (Table 2.6, Figure 2.5). We re-ran the multi-level meta-analysis with a fixed-effect for the categorical publishability ratings and found no difference in mean standardised effect sizes among publishability ratings (Figure A.3).

Out-of-sample Predictions (y_i)

Some models of the influence of reviewer ratings on out-of-sample predictions (y_i) had issues with convergence and singularity of fit (see Table A.8) and those models that converged

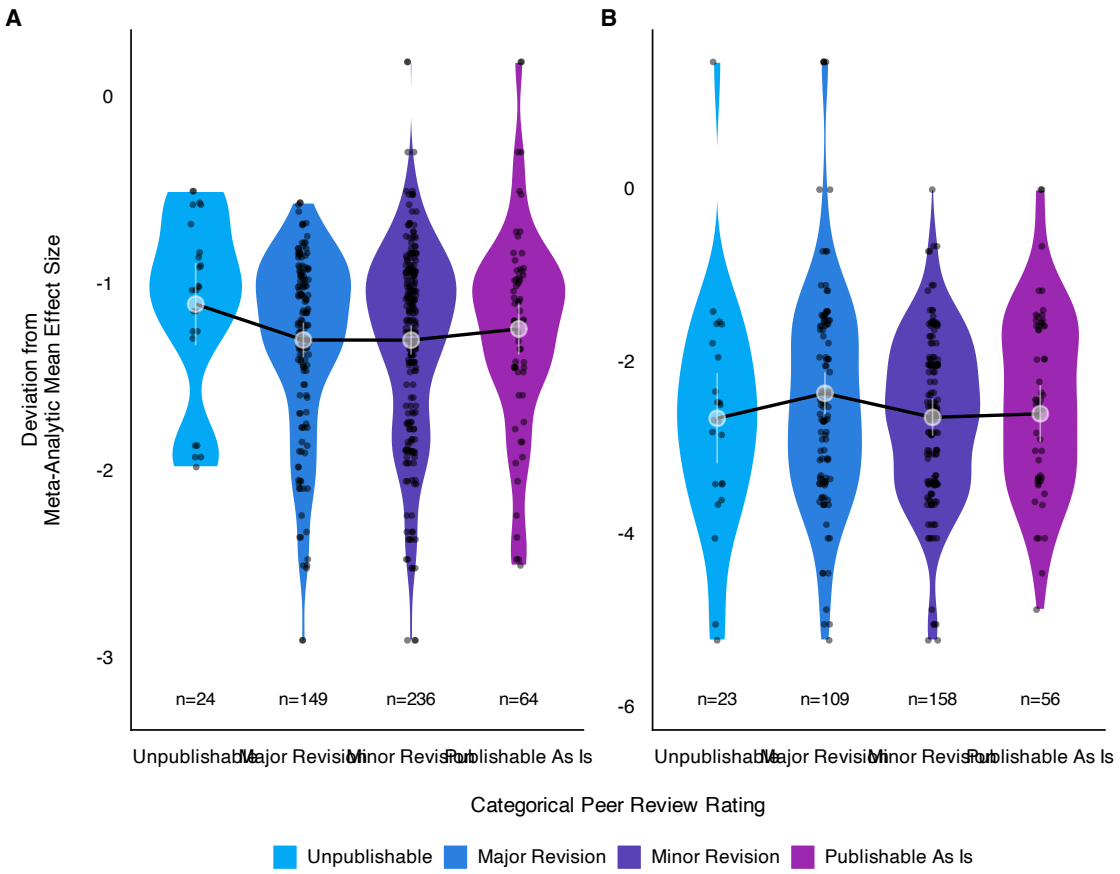


Figure 2.5: *Violin plot of Box-Cox transformed deviation from meta-analytic mean Z_r as a function of categorical peer rating. Grey points for each rating group denote model-estimated marginal mean deviation, and error bars denote 95%CI of the estimate.* ***A*** *Blue tit dataset,* ***B*** Eucalyptus *dataset.*

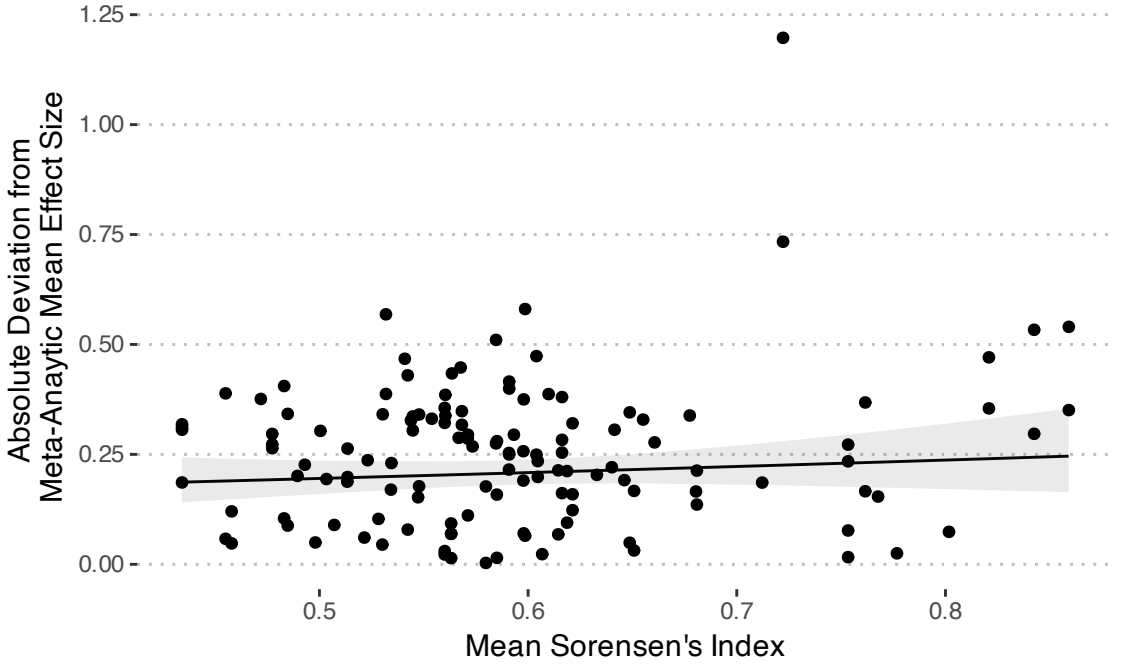
and were not singular showed no strong relationship (Figure A.11, Figure A.12), as with the Zr analyses.

2.3.11 Deviation Scores as Explained by the Distinctiveness of Variables in Each Analysis

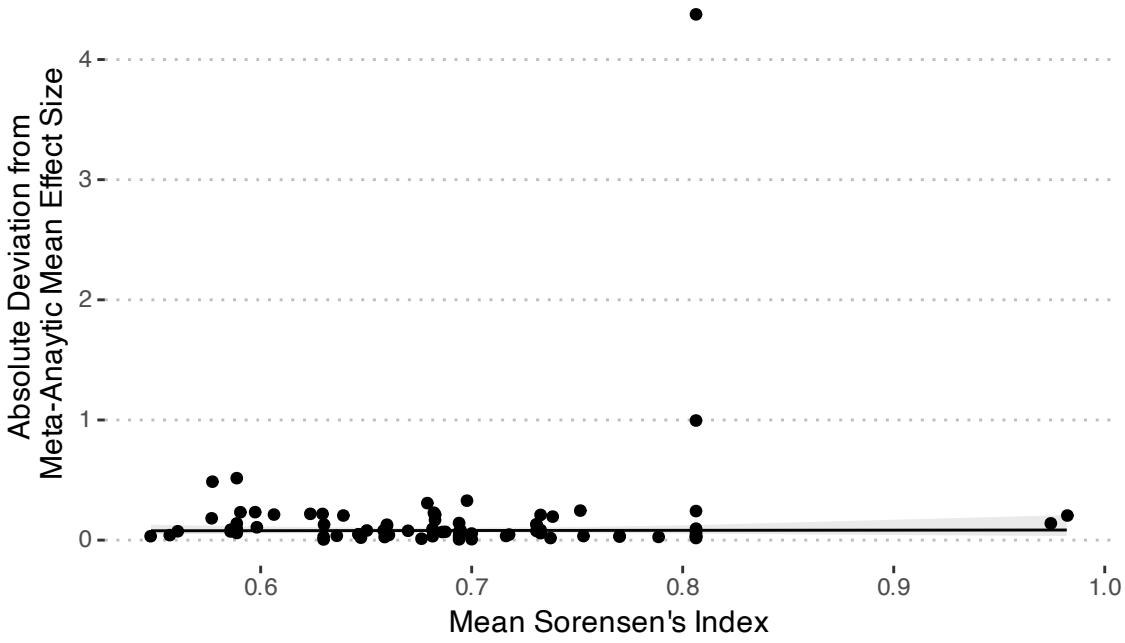
Effect Sizes (Z_r)

We employed Sorensen's index to calculate the distinctiveness of the set of predictor variables used in each model (Figure 2.6). The mean Sorensen's score for blue tit analyses was 0.59 (SD: 0.1, range 0.43-0.86), and for *Eucalyptus* analyses was 0.69 (SD: 0.08, range 0.55-0.98).

We found no meaningful relationship between distinctiveness of variables selected and deviation from the meta-analytic mean (Table 2.6, Figure 2.6) for either blue tit (mean 0.42, 95%CI -0.49,1.34) or *Eucalyptus* effects (mean 0.18, 95%CI -2.78,3.14).



(a) *Blue tit*



(b) Eucalyptus

Figure 2.6: *Fitted model of the Box-Cox transformed deviation score (deviation in effect size from meta-analytic mean) as a function of the mean Sorensen’s index showing distinctiveness of the set of predictor variables. Grey ribbons on predicted values are 95%CI’s.*

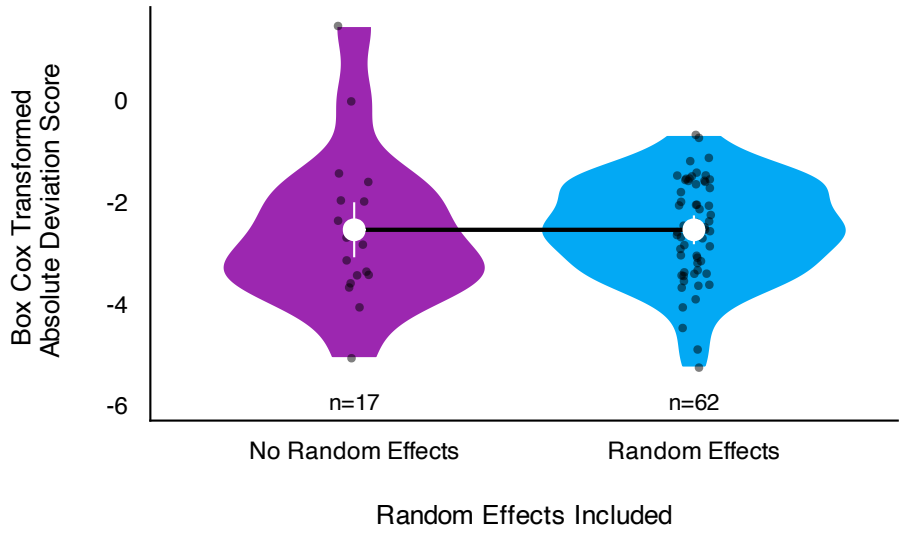


Figure 2.7: *Violin plot of Box-Cox transformed deviation from meta-analytic mean as a function of random-effects inclusion in* Eucalyptus *analyses. White point for each group of analyses denotes model-estimated marginal mean deviation, and error bars denote 95% CI of the estimate.*

Out-of-sample Predictions (y_i)

As with the Z_r estimates, we did not observe any convincing relationships between deviation scores of out-of-sample predictions and Sorensen's index values. Please see supplementary material A.3.4.

2.3.12 Deviation Scores as Explained by the Inclusion of Random Effects

Effect Sizes (Z_r)

There were only three blue tit analyses that did not include random effects, which is below the pre-registered threshold for fitting a model of the Box-Cox transformed deviation from the meta-analytic mean as a function of whether the analysis included random-effects. However, 17 *Eucalyptus* analyses included only fixed effects, which crossed our pre-registered threshold. Consequently, we performed this analysis for the *Eucalyptus* dataset only. There was no relationship between random-effect inclusion and deviation from meta-analytic mean among the *Eucalyptus* analyses (Table 2.6, Figure 2.7).

Out-of-sample Predictions (y_i)

As with the Z_r estimates, we did not examine the possibility of a relationship between the inclusion of random effects and the deviation scores of the blue tit out-of-sample predictions. When we examined the possibility of this relationship for the *Eucalyptus* effects, we found

consistent evidence of somewhat higher Box-Cox-transformed deviation values for models including a random effect, meaning the models including random effects averaged slightly higher deviation from the meta-analytic means (Figure A.14).

2.3.13 Multivariate Analysis Effect Size (Z_r) and Out-of-sample Predictions (y_i)

Like the univariate models, the multivariate models did a poor job of explaining deviations from the meta-analytic mean. Because we pre-registered a multivariate model that contained collinear predictors that produce results which are not readily interpretable, we present these models in the supplement. We also had difficulty with convergence and singularity for multivariate models of out-of-sample (y_i) result, and had to adjust which random effects we included (Table A.13). However, no multivariate analyses of *Eucalyptus* out-of-sample results avoided problems of convergence or singularity, no matter which random effects we included (Table A.13). We therefore present no multivariate *Eucalyptus* y_i models. We present parameter estimates from multivariate Z_r models for both datasets (Table A.11, Table A.12) and from y_i models from the blue tit dataset (Table A.15, Table A.14). We include interpretation of the results from these models in the supplement, but the results do not change the interpretations we present above based on the univariate analyses.

2.4 Discussion

When a large pool of ecologists and evolutionary biologists analyzed the same two datasets to answer the corresponding two research questions, they produced substantially heterogeneous sets of answers. Although the variability in analytical outcomes was high for both datasets, the patterns of this variability differed distinctly between them. For the blue tit dataset, there was nearly continuous variability across a wide range of Z_r values. In contrast, for the *Eucalyptus* dataset, there was less variability across most of the range, but more striking outliers at the tails. Among out-of-sample predictions, there was again almost continuous variation across a wide range (2 SD) among blue tit estimates. For *Eucalyptus*, out-of-sample predictions were also notably variable, with about half the predicted stem count values at <2 but the other half being much larger, and ranging to nearly 40 stems per 15 m x 15 m plot. We investigated several hypotheses for drivers of this variability within datasets, but found little support for any of these. Most notably, even when we excluded analyses that had received one or more poor peer reviews, the heterogeneity in results largely persisted. Regardless of what drives the variability, the existence of such

dramatically heterogeneous results when ecologists and evolutionary biologists seek to answer the same questions with the same data should trigger conversations about how ecologists and evolutionary biologists analyze data and interpret the results of their own analyses and those of others in the literature (e.g., Silberzahn et al., 2018; Auspurg & Brüderl, 2021; Breznau et al., 2022; Simonsohn et al., 2020).

Our observation of substantial heterogeneity due to analytical decisions is consistent with a small earlier study in ecology (Stanton-Geddes et al., 2014) and a growing body of work from the quantitative social sciences (e.g., Silberzahn et al., 2018; Botvinik-Nezer et al., 2020; Breznau et al., 2022; Coretta et al., 2023; Huntington-Klein et al., 2021; Schweinsberg et al., 2021). In these studies, when volunteers from the discipline analyzed the same data, they produced a worryingly diverse set of answers to a pre-set question. This diversity included a wide range of effect sizes, and in most cases, even involved effects in opposite directions. Thus, our result should not be viewed as an anomalous outcome from two particular datasets, but instead as evidence from additional disciplines regarding the heterogeneity that can emerge from analyses of complex datasets to answer questions in probabilistic science. Not only is our major observation consistent with other studies, it is, itself, robust because it derived primarily from simple forest plots that we produced based on a small set of decisions that were mostly registered before data gathering and which conform to widely accepted meta-analytic practices.

Unlike the strong pattern we observed in the forest plots, our other analyses, both registered and *post hoc*, produced either inconsistent patterns, weak patterns, or the absence of patterns. Our registered analyses found that deviations from the meta-analytic mean by individual effect sizes ($\bar{Z}_r$) or the predicted values of the dependent variable ($\bar{y}$) were poorly explained by our hypothesised predictors: peer rating of each analysis team's method section, a measurement of the distinctiveness of the set of predictor variables included in each analysis, or whether the model included random effects. However, in our *post hoc* analyses, we found that dropping analyses identified as unpublishable or in need of major revision by at least one reviewer modestly reduced the observed heterogeneity among the Z_r outcomes, but only for *Eucalyptus* analyses, apparently because this led to the dropping of the major outlier. This limited role for peer review in explaining the variability in our results should be interpreted cautiously because the inter-rater reliability among peer reviewers was extremely low, and at least some analyses that appeared flawed to us were not marked as flawed by reviewers. Thus it seems that the peer reviews we received were of mixed quality,

possibly due to lack of expertise or lack of care on the part of some reviewers. However, the hypothesis that poor quality analyses drove a substantial portion of the heterogeneity we observed was also contradicted by our observation that analysts' self-declared statistical expertise appeared unrelated to heterogeneity. When we retained only analyses from teams including at least one member with high self-declared levels of expertise, heterogeneity among effect sizes remained high. Thus, our results suggest lack of statistical expertise is not the primary factor responsible for the heterogeneity we observed, although further work is merited before rejecting a role for statistical expertise. Besides variability in expertise, it is also possible that the volunteer analysts varied in the effort they invested, and low effort presumably drove at least some heterogeneity in results. However, analysts often submitted thoughtful and extensive code, tables, figures, and textual explanation and interpretations, which is evidence of substantial investment. Further, we are confident that low effort alone is an insufficient explanation for the heterogeneity we observed because we have worked with these datasets ourselves, and we know from experience that there are countless plausible modeling alternatives that can produce a diversity of effects. Additionally, heterogeneity in analytical outcomes differed notably between datasets, and there is no reason to expect that one set of analysts took this project less seriously than the other. Returning to our exploratory analyses, not surprisingly, simply dropping outlier values of Z_r for *Eucalyptus* analyses, which had more extreme outliers, led to less observable heterogeneity in the forest plots, and also reductions in our quantitative measures of heterogeneity. We did not observe a similar effect in the blue tit dataset because that dataset had outliers that were much less extreme and instead had more variability across the core of the distribution.

Our major observations raise two broad questions; why was the variability among results so high, and why did the pattern of variability differ between our two datasets. One important and plausible answer to the first question is that much of the heterogeneity derives from the lack of a precise relationship between the two biological research questions we posed and the data we provided. This lack of a precise relationship between data and question creates many opportunities for different model specifications, and so may inevitably lead to varied analytical outcomes (Auspurg & Brüderl, 2021). However, we believe that the research questions we posed are consistent with the kinds of research question that ecologists and evolutionary biologists typically work from. When designing the two biological research questions, we deliberately sought to represent the level of specificity we typically see in

these disciplines. This level of specificity is evident when we look at the research questions posed by some recent meta-analyses in these fields:

- "how [does] urbanisation impact mean phenotypic values and phenotypic variation … [in] paired urban and non-urban comparisons of avian life-history traits" (Capilla-Lasheras et al., 2022)
- "[what are] the effects of ocean acidification on the crustacean exoskeleton, assessing both exoskeletal ion content (calcium and magnesium) and functional properties (biomechanical resistance and cuticle thickness)" (Siegel et al., 2022)
- "[what is] the extent to which restoration affects both the mean and variability of biodiversity outcomes … [in] terrestrial restoration" (Atkinson et al., 2022)
- "[does] drought stress [have] a negative, positive, or null effect on aphid fitness" (Leybourne et al., 2021)
- "[what is] the influence of nitrogen-fixing trees on soil nitrous oxide emissions" (Kou-Giesbrecht & Menge, 2021)

There is not a single precise answer to any of these questions, nor to the questions we posed to analysts in our study. And this lack of single clear answers will obviously continue to cause uncertainty since ecologists and evolutionary biologists conceive of the different answers from the different statistical models as all being answers to the same general question. A possible response would be a call to avoid these general questions in favor of much more precise alternatives (Auspurg & Brüderl, 2021). However, the research community rewards researchers who pose broad questions (Simons et al., 2017), and so researchers are unlikely to narrow their scope without a change in incentives. Further, we suspect that even if individual studies specified narrow research questions, other scientists would group these more narrow questions into broader categories, for instance in meta-analyses, because it is these broader and more general questions that often interest the research community.

Although variability in statistical outcomes among analysts may be inevitable, our results raise questions about why this variability differed between our two datasets. We are particularly interested in the differences in the distribution of Z_r since the distributions of out-of-sample predictions were on different scales for the two datasets, thus limiting the value of comparisons. The forest plots of Z_r from our two datasets showed distinct

patterns, and these differences are consistent with several alternative hypotheses. The results submitted by analysts of the *Eucalyptus* dataset showed a small average (close to zero) with most estimates also close to zero (±0.2), though about a third far enough above or below zero to cross the traditional threshold of statistical significance. There were a small number of striking outliers that were very far from zero. In contrast, the results submitted by analysts of the blue tit dataset showed an average much further from zero (-0.35) and a much greater spread in the core distribution of estimates across the range of Z_r values (±0.5 from the mean), with few modest outliers. So, why was there more spread in effect sizes (across the estimates that are not outliers) in the blue tit analyses relative to the *Eucalyptus* analyses?

One possible explanation for the lower heterogeneity among most *Eucalyptus* Z_r effects is that weak relationships may limit the opportunities for heterogeneity in analytical outcome. Some evidence for this idea comes from two sets of "many labs" studies in psychology (Klein et al., 2014, 2018). In these studies, many independent lab groups each replicated a large set of studies, including, for each study, the experiment, data collection, and statistical analyses. These studies showed that, when the meta-analytic mean across the replications from different labs was small, there was much less heterogeneity among the outcomes than when the mean effect sizes were large (Klein et al., 2014, 2018). Of course, a weak average effect size would not prevent divergent effects in all circumstances. As we saw with the *Eucalyptus* analyses, taking a radically smaller subset of the data can lead to dramatically divergent effect sizes even when the mean with the full dataset is close to zero.

Our observation that dramatic sub-setting in the *Eucalyptus* dataset was associated with correspondingly dramatic divergence in effect sizes leads us towards another hypothesis to explain the differences in heterogeneity between the *Eucalyptus* and blue tit analysis sets. It may be that when analysts often divide a dataset into subsets, the result will be greater heterogeneity in analytical outcome for that dataset. Although we saw sub-setting associated with dramatic outliers in the *Eucalyptus* dataset, nearly all other analyses of *Eucalyptus* data used close to the same set of 351 samples, and as we saw, these effects did not vary substantially. However, analysts often analyzed only a subset of the blue tit data, and as we observed, sample sizes were much more variable among blue tit effects, and the effects themselves were also much more variable. Important to note here is that subsets of data may differ from each other for biological reasons, but they may also differ due to sampling error. Sampling error is a function of sample size, and sub-samples are, by

definition, smaller samples, and so more subject to variability in effects due to sampling error (Jennions et al., 2013).

Other features of datasets are also plausible candidates for driving heterogeneity in analytical outcomes, including features of covariates. In particular, relationships between covariates and the response variable as well as relationships between covariates and the primary independent variable (collinearity) can strongly influence the modeled relationship between the independent variable of interest and the dependent variable (Dormann et al., 2013; Morrissey & Ruxton, 2018). Therefore, inclusion or exclusion of these covariates can drive heterogeneity in effect sizes (Z_r). Also, as we saw with the two most extreme Z_r values from the blue tit analyses, in multivariate models with collinear predictors, extreme effects can emerge when estimating partial correlation coefficients due to high collinearity, and conclusions can differ dramatically depending on which relationship receives the researcher's attention. Therefore, differences between datasets in the presence of strong and/or collinear covariates could influence the differences in heterogeneity in results among those datasets.

Although it is too early in the many-analyst research program to conclude which analytical decisions or which features of datasets are the most important drivers of heterogeneity in analytical outcomes, we must still grapple with the possibility that analytical outcomes may vary substantially based on the choices we make as analysts. If we assume that, at least sometimes, different analysts will produce dramatically different statistical outcomes, what should we do as ecologists and evolutionary biologists? We review some ideas below.

The easiest path forward after learning about this analytical heterogeneity would be simply to continue with "business as usual", where researchers report results from a small number of statistical models. A case could be made for this path based on our results. For instance, among the blue tit analyses, the precise values of the estimated Z_r effects varied substantially, but the average effect was convincingly different from zero, and a majority of individual effects (84%) were in the same direction. Arguably, many ecologists and evolutionary biologists appear primarily interested in the direction of a given effect and the corresponding *p*-value (Fidler et al., 2006), and so the variability we observed when analyzing the blue tit dataset may not worry these researchers. Similarly, most effects from the *Eucalyptus* analyses were relatively close to zero, and about two-thirds of these effects did not cross the traditional threshold of statistical significance. Therefore, a large proportion of people

analyzing these data would conclude that there was no effect, and this is consistent with what we might conclude from the meta-analysis.

However, we find the counter arguments to "business as usual" to be compelling. For blue tits, there were a substantial minority of calculated effects that would be interpreted by many biologists as indicating the absence of an effect (28%), and there were three traditionally 'significant' effects in the opposite direction to the average. The qualitative conclusions of analysts also reflected substantial variability, with fully half of teams drawing a conclusion distinct from the one we draw from the distribution as a whole. These teams with different conclusion were either uncertain about the negative relationship between competition and nestling growth, or they concluded that effects were mixed or absent. For the *Eucalyptus* analyses, this issue is more concerning. Around two-thirds of effects had confidence intervals overlapping zero, and of the third of analyses with confidence intervals excluding zero, almost half were positive, and the rest were negative. Accordingly, the qualitative conclusions of the *Eucalyptus* teams were spread across the full range of possibilities. But even this striking lack of consensus may be much less of a problem than what could emerge as scientists select which results to publish.

A potentially larger argument against "business as usual" is that it provides the raw material for biasing the literature. When different model specifications readily lead to different results, analysts may be tempted to report the result that appears most interesting, or that is most consistent with expectation (Forstmeier et al., 2017; Gelman & Loken, 2013). There is growing evidence that researchers in ecology and evolutionary biology often report a biased subset of the results they produce (Deressa et al., 2023; Kimmel et al., 2023), and that this bias exaggerates the average size of effects in the published literature between 30 and 150% (Parker & Yang, 2023; Yang et al., 2023). The bias then accumulates in meta-analyses, apparently more than doubling the rate of conclusions of "statistical significance" in published meta-analyses above what would have been found in the absence of bias (Yang et al., 2023). Thus, "business as usual" does not just create noisy results, it helps create systematically misleading results.

If we move away from "business as usual," where do we go? Many obvious options involve multiple analyses per dataset. For instance, there is the traditional robustness or sensitivity check (e.g., Pei et al., 2020; Briga & Verhulst, 2021), in which the researcher presents several alternative versions of an analysis to demonstrate that the result is 'robust' (Lu &

White, 2014). Unfortunately, robustness checks are at risk of the same potential biases of reporting found in other studies (Silberzahn et al., 2018), especially given the relatively few models typically presented. However, these risks could be minimised by running more models and doing so with a pre-registration or registered report. Another option is model averaging. Averages across models often perform well (e.g., Taylor & Taylor, 2023), and in some forms this may be a relatively simple solution. Model averaging, as most often practiced in ecology and evolutionary biology, involves first identifying a small suite of candidate models (see Burnham & Anderson, 2002), then using Akaike weights, based on Akaike's Information Criterion (AIC), to calculate weighted averages for parameter estimates from those models. As with typical robustness checks, the small number of models limits the exploration of specification space, but examining a larger number of models could become the norm. However, there are more concerning limitations. The largest of these limitations is that averaging regression coefficients is problematic when models differ in interaction terms or collinear variables (Cade, 2015). Additionally, weighting by AIC may often be inconsistent with our modelling goals. AIC balances the trade-off between model complexity and predictive ability, but penalizing models for complexity may not be suited for testing hypotheses about causation. So, AIC may often not offer the weight we want to use, and we may also not wish to just generate an average at all. Instead, if we hope to understand an extensive universe of possible modelling outcomes, we could conduct a multiverse analysis, possibly with a specification curve (Simonsohn et al., 2015, 2020). This could mean running hundreds or thousands of models (or more!) to examine the distribution of possible effects, and to see how different model specification choices map onto these effects. However, exploring large areas of specification space may come at the cost of including biologically implausible specifications. Thus, we expect a trade-off, and attempts to limit models to the most biologically plausible may become increasingly difficult in proportion to the number of variables and modeling choices. To make selecting plausible models easier, one could recruit multiple analysts to design one or a few plausible specifications each as with our 'many analyst' study (Silberzahn et al., 2018). An alternative that may be more labor intensive for the primary analyst, but which may lead to a more plausible set of models, could involve hypothesizing about causal pathways with DAGs [directed acyclic graphs; Arif & MacNeil (2023)] to constrain the model set. As with other options outlined above, generating model specifications with DAGs could be partnered with pre-registration to hinder bias from undisclosed data dredging.

Responses to heterogeneity in analysis outcomes need not be limited to simply conducting more analyses, especially if it turns out that analysis quality drives some of the observed heterogeneity. As we noted above, we cannot yet rule out the possibility that insufficient statistical expertise or poor-quality analyses might drive some portion of the heterogeneity we observed. Improving the quality of analyses might be accomplished with a deliberate increase in investment in statistical education. Many ecology and evolutionary biology students learn their statistical practice informally, with many ecology doctoral programs in the USA not requiring a statistics course (Touchon & McCoy, 2016), and no formal courses of any kind included in doctoral degrees in most other countries. In cases where formal investment in statistical education is lacking, informal resources, such as guidelines and checklists, may help researchers avoid common mistakes. However, unless following guidelines or checklists in enforced for publication, the adherence to guidelines is patchy. For example, despite the publication of guidelines for conducting meta-analyses in ecology, the quality of meta-analyses did not improve substantially over time (Koricheva & Gurevitch, 2014). Even in medical research where adherence to guidelines such as the PRISMA standards for systematic reviews and meta-analyses is more highly valued, adherence is often poor (Page & Moher, 2017).

Although we have reviewed a variety of potential responses to the existence of variability in analytical outcomes, we certainly do not wish to imply that this is a comprehensive set of possible responses. Nor do we wish to imply that the opinions we have expressed about these options are correct. Determining how the disciplines of ecology and evolutionary biology should respond to knowledge of the variability in analytical outcome will benefit from the contribution and discussion of ideas from across these disciplines. We look forward to learning from these discussions and to seeing how these disciplines ultimately respond.

2.5 Conclusions

Overall, our results suggest to us that, where there is a diverse set of plausible analysis options, no single analysis should be considered a complete or reliable answer to a research question. Further, because of the evidence that ecologists and evolutionary biologists often present a biased subset of the analyses they conduct (Deressa et al., 2023; Kimmel et al., 2023; Yang et al., 2023), we do not expect that even a collection of different effect sizes from different studies will accurately represent the true distribution of effects (Yang et al.,

2023). Therefore, we believe that an increased level of skepticism of the outcomes of single analyses, or even single meta-analyses, is warranted going forward. We recognize that some researchers have long maintained a healthy level of skepticism of individual studies as part of sound and practical scientific practice, and it is possible that those researchers will be neither surprised nor concerned by our results. However, we doubt that many researchers are sufficiently aware of the potential problems of analytical flexibility to be appropriately skeptical. We hope that our work leads to conversations in ecology, evolutionary biology, and other disciplines about how best to contend with heterogeneity in results that is attributable to analytical decisions.

3 *ManyEcoEvo*: an R package for Analysing Many-analyst Style Data

Preface

In this chapter I present a software paper describing the R package *ManyEcoEvo*, which I used to conduct the analyses in Chapter 2. *ManyEcoEvo* is archived on Zenodo at:

Gould, E., Fraser, H. S., Nakagawa, S., & Parker, T. H. (2026). ManyEcoEvo: an R package for analysing many-analyst style data (v2.8.9000). *Zenodo*. https://doi.org/10.5281/zenodo.10046152

The source code is hosted in a public GitHub repository at github.com/egouldo/ManyEcoEvo. The software is being actively developed in preparation for submission to the Comprehensive R Archive Network (CRAN), prior to submission of the software and manuscript at the Journal of Open Source Software. Consequently, the current version (v2.8.9000) is not the version used to conduct the analyses in Chapter 2 (v2.7.6, https://doi.org/10.5281/zenodo.13690949).

In addition to function documentation accompanying the package, software documentation is also provided as a *pkgdown* (Wickham et al., 2024) site at egouldo.github.io/ManyEcoEvo. The site contains several 'vignettes' that extend the code examples presented in this chapter.

3.1 Summary

Many-analyst studies are an emerging format of crowdsourced scientific analysis (Schweinsberg et al., 2021), in which multiple analysts independently answer the same research question using the same dataset (Silberzahn et al., 2018; Voracek et al., 2019). Several high-profile studies demonstrate that results and conclusions can be highly contingent

on subjective but defensible analytical decisions (Botvinik-Nezer et al., 2020; Breznau et al., 2021; Coretta et al., 2023; Hoogeveen et al., 2023; Huntington-Klein et al., 2021; Silberzahn et al., 2018), arising from disagreement about modelling strategy or variation in data preparation (Simonsohn et al., 2020; Voracek et al., 2019). Similarly to meta-analysis, many-analyst studies treat analytical variation as the object of study, and produce rich metadata for each analysis — including model type, transformations, and variable selection — that can be used to characterise analytical uncertainty. Unlike meta-analysis, the source of variation under study is analytical rather than arising from sampling or measurement error between independent datasets.

`ManyEcoEvo` is an R package that provides a structured, reproducible workflow for conducting the statistical analyses characteristic of many-analyst studies. Despite being named after a particular 'Many' study, the package implements a pipeline for use in any future 'Many' that: transforms and standardises heterogeneous analyst estimates onto a common scale; fits multilevel meta-analytic models across multiple analyst submissions; computes each analyst's deviation from the meta-analytic mean; and fits a series of univariate regression models examining how analytical choices relate to deviation from the consensus estimate among teams. The analysis workflow is built around a nested tibble structure that allows the full sequence of operations to be mapped systematically over multiple datasets and over systematically varied analytical subsets of those datasets — for example, subsets defined by peer-review quality ratings, variable collinearity, or analyst expertise. This architecture makes repeated application of the analysis pipeline across many dataset variants a tractable problem.

Although `ManyEcoEvo` was developed for the 'Same Data, Different Analysts' study in ecology and evolutionary biology (Chapter 2), its workflow generalises to computationally analogous problems. Collaborative big-team open science projects (e.g. 'ManyBirds,' ManyBirds Project et al., 2025; 'ManyBabies,' Bergmann et al., 2025; McMillan et al., 2026; 'ManyPrimates,' Bohn et al., 2024) and coordinated distributed experiments (e.g. 'NutNet,' Borer et al., 2014; 'DragNet,' Borer & Seabloom, 2025) produce multiple datasets that pose similar data-manipulation and analysis challenges. The package also generalises to sensitivity analyses in conventional meta-analyses, where the primary analysis is repeated across varied inclusion criteria or data subsets.

In addition to the workflow infrastructure, `ManyEcoEvo` ships the curated, analysis-ready datasets from Chapter 2 as package data objects. These include analyst effect-size estimates and metadata, out-of-sample predictions as an alternative outcome variable, peer-review ratings, and variable-use data for two research questions — one from evolutionary biology (blue tit nestling growth) and one from conservation ecology (*Eucalyptus* seedling recruitment). The packaged data enables straightforward reproduction and re-analysis of the published results.

3.1.1 Statement of Need

Conducting a many-analyst study requires solving a class of data-management and analytical problems that existing meta-analytic software does not address. The primary challenge is heterogeneity: analyst submissions span a wide variety of modelling approaches, response variable transformations, and link functions, all of which must be identified, back-transformed, and standardised onto a common scale before any synthesis is possible. Tracing each analyst's reported transformation, applying the appropriate back-transformation, and propagating the result through to a standardised effect size or out-of-sample prediction, is extremely time-consuming and error-prone at scale.

A second challenge is systematic repetition. Many-analyst studies - especially like Chapter 2 where different analysis questions were posed - require the full many-analyst pipeline to be repeated across multiple analytical subsets. For example, on subsets; with analyses that were rated as low quality by peer reviewers, or that contained highly collinear predictors; with extreme outliers removed; that were restricted to expert analysts. Without dedicated infrastructure, this repetition is implemented as bespoke, study-specific scripts: copies of the primary analysis, manually adapted for each subset. This approach makes it difficult to reason about what differs between subsets and what does not, obscures whether observed differences in results across subsets are attributable to the subset definition or to incidental differences in the code, and makes it hard to verify that the same analytical protocol has been applied consistently. It also makes the analysis difficult to reproduce, audit, or adapt to new data.

The target audience for `ManyEcoEvo` is therefore researchers conducting or re-analysing many-analyst studies, as well as researchers whose work involves the systematic repetition of a meta-analytic pipeline across multiple datasets or analytically defined data subsets.

This includes teams running collaborative big-team open science projects where data are collected across multiple experiments, locations or analyst groups. A secondary audience is conventional meta-analysts for whom sensitivity analyses, which repeat the primary meta-analysis under varied inclusion criteria, are a routine but currently ad hoc component of their workflow.

`ManyEcoEvo` addresses these challenges by providing a reusable, programmatic workflow that treats outcome-variable harmonisation and standardisation, subset generation, and pipeline repetition as first-class operations. Standardisation and back-transformation of heterogeneous analyst estimates are handled by a family of functions that identify and apply the appropriate transformation for each analysis, making the process explicit, auditable, and reproducible. Subset generation is implemented as row-expansion of a nested tibble, so that the full downstream pipeline — meta-analysis, computation and transformation of deviation scores, and univariate regressions — propagates automatically through all dataset and subset combinations without modification to the pipeline itself. This architecture makes it straightforward to verify that the same analytical protocol has been applied consistently across all subsets, and to isolate the effect of each subset definition on results. Together, these properties support FAIR (Wilkinson et al., 2016) and reproducible research practice in a study format that is otherwise difficult to implement in a principled way.

3.2 State of the Field

The R ecosystem for meta-analysis is mature and well-characterised. Lortie & Filazzola (2020) provide a structured contrast of the two most widely used packages in ecology and evolutionary biology: meta (Balduzzi et al., 2019), a direct and intuitive choice for rapid implementation of general meta-analytic statistics, and `metafor` (Viechtbauer, 2010), a more comprehensive package suited to complex multilevel models. Both provide heterogeneity estimation, visualisation tools, and publication bias diagnostics, and both are widely used across ecology, evolution, and the social sciences. For visualisation, `orchaRd` (Nakagawa, Lagisz, et al., 2023) extends `metafor` with orchard plots suited to ecological effect-size data, and `metaviz` (Kossmeier et al., 2020) provides additional forest and funnel plot variants. These packages form the analytical foundation on which `ManyEcoEvo` builds: the multilevel meta-regression at the core of the package is implemented via `metafor::rma.mv()`, and forest plots are produced using conventions established by

`orchaRd` and `metaviz`. However, all of these packages share a common design assumption — that the unit of analysis is a single meta-analytic dataset, and that the researcher's task is to fit, interrogate, and visualise one model. None provide infrastructure for the upstream data-wrangling challenges specific to many-analyst data, nor for the systematic generation and propagation of analytical subsets through a complete pipeline. Contributing the `ManyEcoEvo` workflow to any of these packages would require introducing a fundamentally different data model and pipeline architecture that sits outside their intended scope.

A second category of potentially relevant tools comes from multiverse analysis and specification curve analysis. The `multiverse` package (Sarma et al., 2023) provides an embedded domain-specific language for declaring alternative analysis paths within R and R Notebooks, implementing the multiverse analysis philosophy of Steegen et al. (2016). `specr` (Masur & Scharkow, 2020) provides utilities for conducting and visualising specification curve analyses (Simonsohn et al., 2020), allowing users to set up, run, evaluate, and plot the full space of analytical specifications for a single dataset. `Boba` (Liu et al., 2021) extends this further with an integrated domain-specific language and interactive visualisation system for authoring and interpreting multiverse analyses across multiple programming languages. These tools share with `ManyEcoEvo` the goal of characterising how analytical variation affects results, but they operate in a fundamentally different setting: they generate variation in analysis outcomes programmatically and combinatorially from a single analyst's declared decision space, rather than synthesising analytic variation across many analysts or teams. The conceptual framing and associated analysis of variation is also distinct: multiverse and specification curve tools ask "what would have happened if this analyst had made different choices?", whereas `ManyEcoEvo` asks "given that many analysts made different choices, what can we infer about the distribution and drivers of those differences?" Consequently, they do not address the resolution of incompatibility between analyst estimates arising from different model specifications, transformations, and link functions prior to meta-analytic synthesis. Contributing the `ManyEcoEvo` pipeline to any of these tools would require introducing a meta-analytic synthesis layer and a model for resolving heterogeneous-input data that are outside their design scope entirely.

The closest precedents for the `ManyEcoEvo` approach are the study-specific analysis repositories published alongside individual many-analyst studies, such as those accompanying the 'ManyBabies' (Bergmann et al., 2025) and 'ManyPrimates' (Bohn et al., 2024) projects. These repositories demonstrate that the analytical problems `ManyEcoEvo` addresses recur

across disciplines. However, their code is written for a single study and dataset, with no abstraction or reuse intended. Contributing to these repositories would produce more study-specific code rather than a reusable, generalisable tool. `ManyEcoEvo` was therefore developed as a standalone package because no existing tool addresses the combination of resolution of heterogeneity in outcome variables, nested data management, systematic subset generation, and reproducible pipeline repetition that many-analyst studies require.

3.3 Research Impact

`ManyEcoEvo` was developed to support the analysis reported in Chapter 2, a large-scale many-analyst study in ecology and evolutionary biology in which 174 analyst teams independently investigated the same two research questions. That paper has accumulated 54 citations, over 40,000 accesses, and an Altmetric score of 347 since publication, indicating substantial engagement from the research community. The packaged datasets and workflow have been independently used by Knief & Forstmeier (2025), who conducted a targeted multiverse re-analysis of the blue tit dataset from Chapter 2, demonstrating that the curated data objects shipped with `ManyEcoEvo` are sufficient to support independent re-analysis and extension of the published results.

The package is archived on Zenodo with a citable DOI, and has accumulated 670 views and 150 downloads across all versions. The GitHub repository (`https://github.com/egouldo/ManyEcoEvo`) has received 264 clones from 91 unique cloners in the 14 days prior to submission, suggesting active and ongoing interest from the research community. Package documentation is maintained at a dedicated `pkgdown` site (Wickham et al. (2024), `https://egouldo.github.io/ManyEcoEvo/`), and the repository follows community standards for R package development (Wickham & Bryan, 2023), including a ticketed issue tracker, open source contribution workflow, and contributor guidelines accompanied by a code of conduct. These considerations collectively indicate that the package is ready for community use and external contribution.

3.4 Software Design

`ManyEcoEvo` is built around a nested tibble as its central data structure, in which each row represents a dataset or analytical subset, and the analyst-level data for that row is stored in list-columns rather than as flat rows. This design was chosen deliberately because it

maps naturally onto the problem structure of many-analyst studies: the same sequence of operations must be applied repeatedly and consistently across multiple datasets and multiple variants of those datasets. Many-analyst data also has an inherently hierarchical structure — a single research question may receive multiple analyses from the same team, and each analysis may be reviewed by multiple reviewers — that requires grouped operations to handle correctly. By representing each dataset or subset as a row with analyst-level data stored in a list-column, grouping and iteration are expressed cleanly and explicitly: the full pipeline can be applied as row-wise transformations using `purrr::map()` variants from the tidyverse (Wickham et al., 2019), and grouped operations within a dataset are handled naturally by the nested structure without requiring fragile index-based bookkeeping. This makes the pipeline both concise and legible: the structure of the code mirrors the structure of the analysis, and adding a new dataset or subset requires only an additional row rather than a modification to the pipeline itself.

A flat data frame approach was considered and rejected. While flat data frames are straightforward for simple operations, they become unwieldy when joining heterogeneous nested objects — such as data on the distinctiveness of analysis variables, peer-review ratings, and back-transformed point estimates — across datasets for downstream plotting and analysis. They are also poorly suited to the programmatic generation of multiple subset variants, where each variant requires its own correctly filtered copy of several related data objects. Managing these relationships in a flat structure requires either duplicating data or maintaining fragile index-based joins, both of which are error-prone at the scale of a many-analyst study. The nested tibble approach keeps related data objects co-located in the same row, making the relationships between them explicit and reducing the risk of misalignment.

The pipeline also makes a deliberate distinction between modular analysis-level functions, such as `est_to_zr()`, `calculate_deviation_score()`, and `box_cox_transform()`, and wrapper functions like `meta_analyse_datasets()`, which coordinate the full sequence for a given dataset. This two-level architecture means that users who wish to apply only part of the pipeline, substitute their own standardisation method, or adapt the workflow to a different study design can do so by calling the modular functions directly, without being constrained by the wrapper. The wrapper functions exist for convenience and reproducibility, not as the only entry point.

Back-transformation of heterogeneous analyst estimates to the response variable scale is handled via simulation: for each analyst, a distribution of estimates is generated by sampling from a normal distribution parameterised by the reported estimate and standard error, the appropriate inverse transformation is applied to the full distribution, and the mean and standard deviation of the transformed distribution are returned. The simulation approach was preferred because it is general, handling any transformation for which an inverse exists, and because it makes the back-transformation procedure transparent and independently verifiable. `ManyEcoEvo` provides wrapper functions that dispatch the analysts' estimates to the correct back-transformation function automatically when information about the analyst's choice of link-function and response variable scales are present in the dataset.

The relationship between the software and the associated publication warrants brief mention. Version 2.7.6 of `ManyEcoEvo`, archived on Zenodo with a versioned DOI (https://doi.org/10.5281/zenodo.13690949) corresponds exactly to the software used to produce the results reported in Chapter 2. The current development version (2.8.0.9000) extends this release and is distinguished from it by both the Zenodo version DOI (https://doi.org/10.5281/zenodo.10046152) and the GitHub release (https://github.com/egouldo/ManyAnalysts/releases). Users wishing to exactly reproduce the results of Chapter 2 should install the archived version following the README instructions (https://github.com/egouldo/ManyAnalysts/blob/main/README.md); users wishing to apply the workflow to new data should use the current release.

Package dependencies for regenerating the `ManyEcoEvo` data are managed with `renv` (Ushey & Wickham, 2024), which records the full set of package versions in a project-specific library, ensuring that the computational environment can be restored exactly on a clean install. The analysis pipeline is implemented using the `targets` package (Landau, 2021), which makes the dependency structure between pipeline steps explicit and provides caching so that only outdated steps are re-executed when inputs change. Data integrity throughout the pipeline is enforced using `pointblank` (Iannone et al., 2025), which performs programmatic validation checks on data structures at key steps, catching malformed inputs early and making data quality expectations explicit and auditable. A formal unit test suite is implemented using `testthat` (Wickham, 2011). Full instructions for a clean install with regeneration of the package data are provided in the package README at https://github.com/egouldo/ManyEcoEvo.

3.5 Software Demonstration

Below I demonstrate how `ManyEcoEvo` can be used to build a many-analyst data pipeline using its nested tibble approach and multi-level wrapper functions. I illustrate this example with the 'Same Data, Different Analysts' dataset (Chapter 2), beginning with the simplest case of meta-analysing a single dataset, before scaling up to map meta-analyses over different datasets, and data subsets.

3.5.1 Dataset Description

The `ManyEcoEvo` dataset contains analyst responses for two different research questions, each with their own analysis dataset: one examining blue tit nestling growth (evolutionary biology) and another examining eucalyptus seedling recruitment (conservation ecology, Chapter 2).

```
library(ManyEcoEvo)
library(tidyverse)
library(metafor)
set.seed(1234)
```

```
# Load dataset
data("ManyEcoEvo")
# Examine the basic structure
ManyEcoEvo
```

```
#> # A tibble: 2 x 4
#>   dataset    data                diversity_data      estimate_type
#>   <chr>      <list>              <named list>        <chr>
#> 1 blue tit   <tibble [174 x 38]> <tibble [174 x 54]> Zr
#> 2 eucalyptus <tibble [128 x 38]> <tibble [128 x 61]> Zr
```

```
glimpse(ManyEcoEvo)
```

```
#> Rows: 2
#> Columns: 4
#> $ dataset        <chr> "blue tit", "eucalyptus"
#> $ data           <list> [<tbl_df[174 x 38]>], [<tbl_df[128 x 38]>]
#> $ diversity_data <named list> [<tbl_df[174 x 54]>], [<tbl_df[128 x 61]>]
#> $ estimate_type  <chr> "Zr", "Zr"
```

The data uses a nested data structure (Wickham, Çetinkaya-Rundel, et al., 2023, Chapter 23; Wickham et al., 2025, https://tidyr.tidyverse.org/articles/nest.html), with a row for each analysis dataset, and two list-columns:

- `data`, which contains the effect-size estimates and analyst metadata as tibbles for each analysis dataset, and
- `diversity_data`, which contains information about variables used by each analysis team, also as tibbles.

For illustration purposes, only a subset of the blue tit dataset is used below.

```
# Extract only necessary columns for illustration
key_cols <- c(
  "id_col",
  "TeamIdentifier",
  "beta_estimate",
  "beta_SE",
  "sample_size",
  "adjusted_df",
  "dataset",
  "mixed_model"
)

analysis_data <- ManyEcoEvo %>%
  mutate(
    data = map(
      data,
      ~ .x %>%
        filter(if_all(all_of(key_cols), ~ !is.na(.))) %>%
        slice_sample(n = 25) %>%
        arrange(TeamIdentifier, analysis_id)
    ),
    diversity_data = map2(
      diversity_data,
      data,
      ~ semi_join(.x, .y, by = "id_col")
    )
  )

effect_sizes <- analysis_data %>%
```

```
  pluck("data", 1) %>%
  select(all_of(key_cols))


head(effect_sizes)
```

```
#> # A tibble: 6 x 8
#>   id_col    TeamIdentifier beta_estimate beta_SE sample_size adjusted_df dataset
#>   <chr>     <chr>                  <dbl>   <dbl>       <dbl>       <dbl> <chr>
#> 1 Angasto-~ Angasto              -0.298   0.0525        3720       548.  blue t~
#> 2 Angasto-~ Angasto               0.0460  0.0866        3720       495.  blue t~
#> 3 Angasto-~ Angasto              -0.211   0.0801        2626       295.  blue t~
#> 4 Antwerp-~ Antwerp              -0.0147  0.0887        3720       871.  blue t~
#> 5 Appin-2-~ Appin                 0.0474  0.0419         599        89.2 blue t~
#> 6 Aramac-1~ Aramac               -0.0222  0.108         3720       513.  blue t~
#> # i 1 more variable: mixed_model <dbl>
```

Each row represents a unique analysis (`id_col`), potentially with multiple analyses per team (`TeamIdentifier`).

3.5.2 Calculating meta-analytic results from raw estimates with the blue tit dataset

Here, I introduce analysis-level functions in `ManyEcoEvo` demonstrating a typical analysis pipeline for a single analysis dataset. Each of these functions is modular and usable independently of the package, so users can substitute their own standardisation method, their own model-fitting function, or their own deviation score approach.

Data preparation for meta-analysis

The first step in any meta-analysis is ensuring that the estimands of interest – in this case, effect sizes as correlation coefficients – are on the same response scale to enable comparison of effect-sizes from heterogenous analyses. The meta-analyses in the `ManyEcoEvo` study are performed on Fishers' Z transformed correlation coefficients Z_r with the function `est_to_zr()` (Equation 2.3, Equation 2.4), which is called by `standardise_response()`, adding columns of standardised estimates `Zr` and their variance `VZr`:

```
# label: glimpse-effect-sizes
standardised_effect_sizes <- effect_sizes %>%
  standardise_response(estimate_type = "Zr")
standardised_effect_sizes %>%
  glimpse()
```

```
#> Rows: 25
#> Columns: 10
#> $ id_col         <chr> "Angasto-1-1-2", "Angasto-2-2-1", "Angasto-3-3-2", "Ant~
#> $ TeamIdentifier <chr> "Angasto", "Angasto", "Angasto", "Antwerp", "Appin", "A~
#> $ beta_estimate  <dbl> -0.29782521, 0.04597628, -0.21100522, -0.01470000, 0.04~
#> $ beta_SE        <dbl> 0.052517300, 0.086562620, 0.080133580, 0.088700000, 0.0~
#> $ sample_size    <dbl> 3720, 3720, 2626, 3720, 599, 3720, 3720, 2629, 2629, 25~
#> $ adjusted_df    <dbl> 547.81510, 494.94250, 294.82840, 871.33200, 89.21005, 5~
#> $ dataset        <chr> "blue tit", "blue tit", "blue tit", "blue tit", "blue t~
#> $ mixed_model    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
#> $ Zr             <dbl> -0.239983502, 0.023871781, -0.152758804, -0.005614352, ~
#> $ VZr            <dbl> 0.0018254334, 0.0020204367, 0.0033918035, 0.0011476682,~
```

Now that the analysts' estimates have been standardised, the meta-analysis can progress.

Meta-analysis and univariate analyses

Following Chapter 2, I fit a metaregression to the standardised effect sizes, using the `metafor` package, with nested random effects (Equation 2.5):

```
standardised_effect_sizes <- standardised_effect_sizes %>%
  mutate(study_id = id_col)
ma_model <- fit_MA_mv(
  effects_analysis = standardised_effect_sizes,
  outcome_colname = "Zr",
  outcome_SE_colname = "VZr",
  estimate_type = "Zr"
)
summary(ma_model)
```

```
#>
#> Multivariate Meta-Analysis Model (k = 25; method: REML)
#>
#>   logLik  Deviance       AIC       BIC      AICc
#>  -0.3937    0.7875    6.7875   10.3216    7.9875
#>
```

```
#> Variance Components:
#>
#>              estim    sqrt  nlvls  fixed                    factor
#> sigma^2.1  0.0479  0.2190     18     no           TeamIdentifier
#> sigma^2.2  0.0217  0.1474     25     no  TeamIdentifier/study_id
#>
#> Test for Heterogeneity:
#> Q(df = 24) = 558.6450, p-val < .0001
#>
#> Model Results:
#>
#> estimate      se     zval    pval    ci.lb    ci.ub
#>  -0.3091  0.0616  -5.0191  <.0001  -0.4298  -0.1884  ***
#>
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculate Inputs for Univariate Analyses In Chapter 2, I examined how different analysis features influenced the deviation of estimated effect sizes from the meta-analytic mean effect size, including the distinctiveness of variables, peer-review ratings (as a numerical and categorical score), and whether the analyses included random effects (Section 2.3.9). First, the outcome variable must be created by calculating the deviation of Z_r from the meta-analytic mean for each analysis `id_col`. The deviation scores are then Box-Cox transformed:

```
effect_size_data <- standardised_effect_sizes %>%
  calculate_deviation_score(
    meta_analytic_model = ma_model,
    outcome_colname = "Zr"
  ) %>%
  box_cox_transform(dataset = "blue tit", outcome_SE_colname = "VZr")
```

Regression on Distinctiveness of Analysis Variables Below I fit the univariate regression on the distinctiveness of analysis variables, measured as the mean Sorensen's pairwise dissimilarity index (Equation 2.8). This approach treats each analysts' variable set as a presence/absence matrix (stored in `diversity_data`) to compute a mean pairwise dissimilarity score per analysis:

```
# Extract diversity data
diversity_data <- analysis_data %>%
  pluck("diversity_data", 1)
mean_sorensen_data <-
  diversity_data %>%
  calculate_sorensen_diversity_index(.id = "id_col")
mean_sorensen_data
```

```
#> # A tibble: 25 x 3
#> # Rowwise:
#>    id_col       mean_diversity_index num_variables
#>    <chr>                       <dbl>         <dbl>
#>  1 Bega-2-2-1                  0.523             8
#>  2 Bruc-1-1-1                  0.423            12
#>  3 Appin-2-2-1                 0.634             3
#>  4 Bargo-3-3-1                 0.456            15
#>  5 Aramac-1-1-1                0.372             9
#>  6 Aramac-1-1-2                0.372             9
#>  7 Ararat-2-2-1                0.395            10
#>  8 Ararat-1-1-1                0.351            10
#>  9 August-2-2-1                0.422             5
#> 10 Barham-2-2-1                0.540            10
#> # i 15 more rows
```

The mean Sorensen's index data is then joined with the deviation scores to fit a univariate `glm` model of the deviation scores on mean Sorensen's diversity index (Equation 2.9). The model is fitted using `parsnip` package and is therefore returned as a `parsnip` object (Kuhn & Vaughan, 2024).

```
effect_size_data <- effect_size_data %>%
  left_join(mean_sorensen_data, by = join_by(id_col))

diversity_model <- fit_sorensen_glm(effect_size_data)
diversity_model # View model summary
```

```
#> parsnip model object
#>
#>
#> Call:  stats::glm(formula = ..y ~ ., family = stats::gaussian, data = data)
#>
#> Coefficients:
```

```
#>          (Intercept)  mean_diversity_index
#>              -0.7985                0.1887
#>
#> Degrees of Freedom: 24 Total (i.e. Null);  23 Residual
#> Null Deviance:       0.2857
#> Residual Deviance: 0.2793    AIC: -35.41
```

Visualise Results ManyEcoEvo includes several data extraction and plotting functions for visualising the meta-analysis and univariate models. Including a forest plot, shown for the blue tit meta-analyses:

```
get_forest_plot_data(ma_model) %>%
  plot_forest(intercept = TRUE)
```

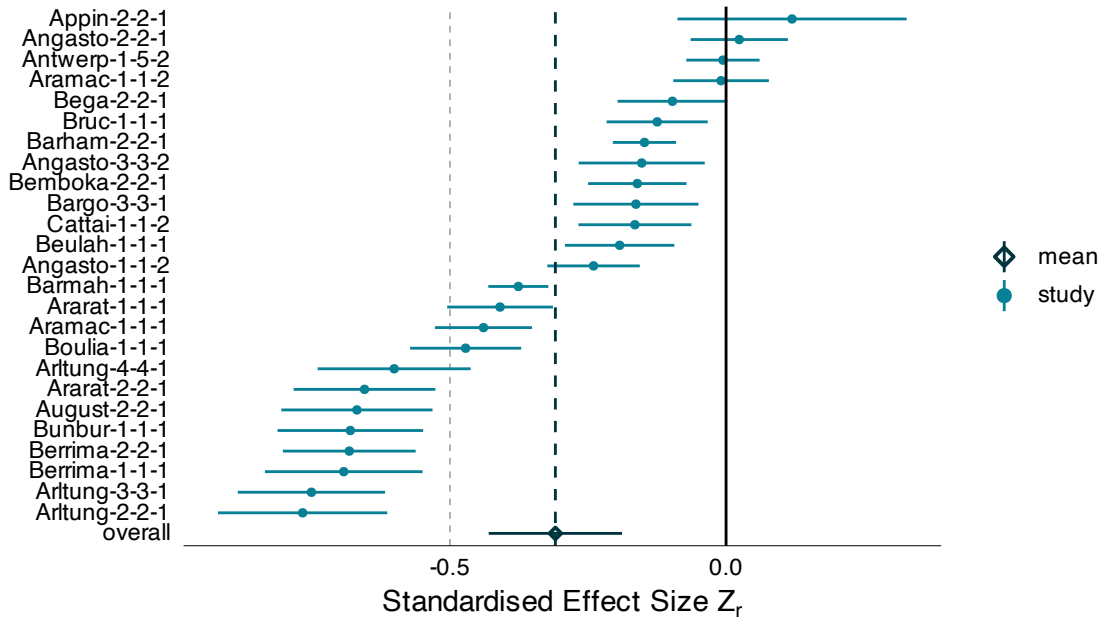


Figure 3.1: *Forest plots of meta-analytic estimated standardised effect sizes (Z_r, blue circles) and their 95% confidence intervals for each effect size included in the meta-analysis model. Results are shown for a subset of the complete blue tit dataset from Chapter 2. Meta-analytic mean effect size is shown as diamond.*

Below I plot the marginal effects for the model of deviation in analysis scores on the distinctiveness of variables, back-transforming Box-Cox transformed response variables for easier visual interpretation:

```
p_sorensen <- plot_effects_diversity(
  mod = diversity_model,
  df = effect_size_data,
  back_transform = TRUE
```

```
)
p_sorensen +
  theme_bw()
```

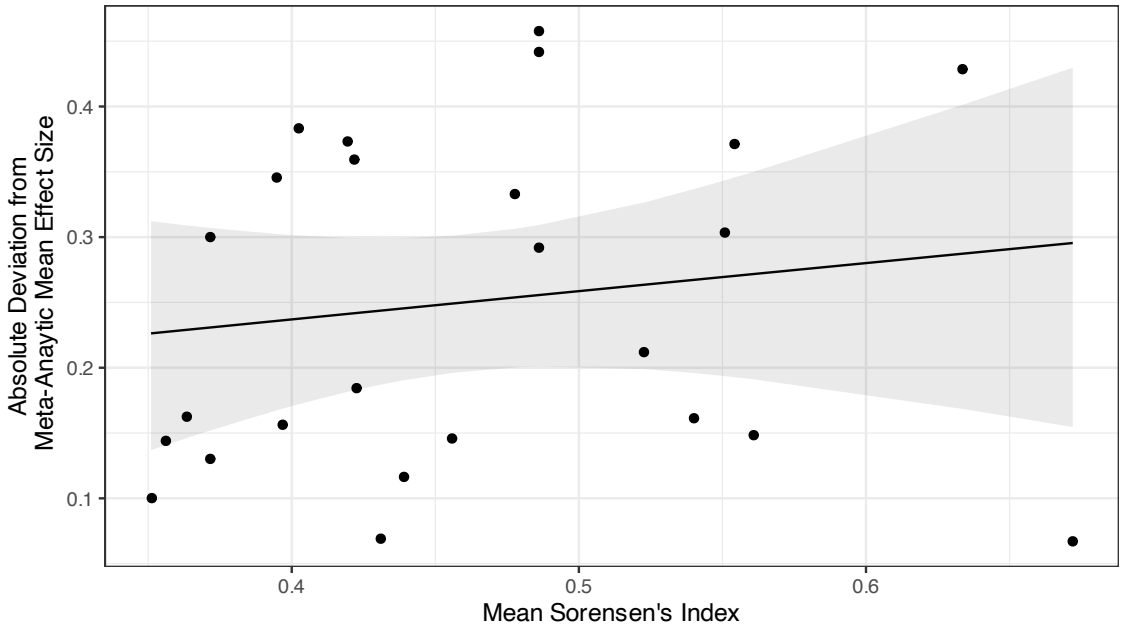


Figure 3.2: *Fitted model of the Box-Cox transformed deviation score (deviation in the effect size from the meta-analytic mean) as a function of the mean Sorensen's index showing distinctiveness of the set of predictor variables, for a subset of the blue tit dataset from Chapter 2. Grey ribbons indicate predicted values are 95% CI's. Deviation scores have been back-transformed to the original scale.*

Since the returned plot is a `ggplot2` object (Wickham, 2016) the aesthetics can be easily customised, as illustrated above by changing the plot theme.

3.6 Scaling up: Systematic Analysis Across Datasets and Analytical Subsets

To demonstrate the logic of the analysis pipeline, each step was introduced individually above, moving from standardising estimates, fitting the meta-analytic model, computing deviation scores, to fitting univariate regressions and visualising results. To simplify scaling up to multiple analyses across different datasets and subsets of those datasets, `ManyEcoEvo` provides a single wrapper function to coordinate each step, `meta_analyse_datasets()`, which executes the full sequence in one call. The function is introduced and applied to a single dataset first, before illustrating how the same call extends to multiple datasets and analytical subsets.

3.6.1 Preparing to map analyses over datasets with the nested tibble

The pipeline is built around a nested tibble — a data frame where each row represents a dataset (e.g., "blue tit" or "eucalyptus"), and the analyst-level data for that dataset is stored in list-columns rather than as flat rows. The key list-columns are shown below:

```
# The top-level object: one row per analysis dataset
ManyEcoEvo
```

```
#> # A tibble: 2 x 4
#>   dataset    data                 diversity_data      estimate_type
#>   <chr>      <list>               <named list>        <chr>
#> 1 blue tit   <tibble [174 x 38]> <tibble [174 x 54]> Zr
#> 2 eucalyptus <tibble [128 x 38]> <tibble [128 x 61]> Zr
```

Each element of the column `data` is itself a tibble of analyst submissions, for each `dataset`. The nesting means that all downstream operations can be expressed as row-wise transformations using `purrr::map()` variants — the pipeline never needs to unpack and repack the full flat data frame. `diversity_data` stores a tibble of variable-use data for the Sorensen index computation, while `estimate_type` describes the outcome variable. An equivalent dataset for point predictions, rather than effect-sizes, is stored separately in `ManyEcoEvo_yi`.

3.6.2 Coordinating the Full Pipeline with `meta_analyse_datasets()`

`meta_analyse_datasets()` accepts the nested tibble produced by `prepare_response_variables()` and `compute_MA_inputs()` and, for each row, executes in sequence:

1. Removal of infinite and missing values from the outcome columns.
2. Fitting of the multilevel meta-analytic model via `fit_MA_mv()`, a wrapper function around `metafor::rma.mv()` function from the `metafor` package (Viechtbauer, 2010).
3. Computation of each analyst's deviation from the meta-analytic mean via `calculate_deviation_score()`.
4. Box–Cox transformation of the deviation scores via `box_cox_transform()`.
5. Fitting of the univariate regression models.

Each result is returned as a new list-column in the same nested tibble. This call is now applied to the *Eucalyptus* dataset:

```
single_dataset_results <-
  analysis_data %>%
  filter(dataset == "eucalyptus") %>%
  prepare_response_variables(
    estimate_type = "Zr",
    dataset_standardise = "eucalyptus"
  ) %>%
  compute_MA_inputs(estimate_type = "Zr") %>%
  meta_analyse_datasets(
    outcome_variable = "Zr",
    outcome_SE = "VZr"
  )

# One row, with all model outputs as list-columns
single_dataset_results %>%
  select(dataset, MA_mod, sorensen_glm, box_cox_rating_cont, box_cox_rating_cat)
```

```
#> # A tibble: 1 x 5
#>   dataset    MA_mod   sorensen_glm box_cox_rating_cont box_cox_rating_cat
#>   <chr>      <list>   <list>       <list>              <list>
#> 1 eucalyptus <rma.mv> <fit[+]>     <lmerMod>           <lmerMod>
```

This is equivalent to the manual sequence demonstrated in the previous section (Section 3.5.2), but expressed as a single pipeline call whose structure does not change as the number of datasets or subsets grows.

3.6.3 Repeating the Analysis Across Multiple Datasets

Extending the pipeline to both datasets requires no changes to the pipeline itself — only the input object changes. Because `meta_analyse_datasets()` iterates over rows of the nested tibble, passing a two-row object (one row per dataset) produces two complete sets of results:

```
meta_analysis_results
```

```
#> # A tibble: 2 x 13
#>   dataset    data     diversity_data     diversity_indices   effects_analysis
#>   <chr>      <list>   <named list>       <named list>        <list>
```

```
#> 1 blue tit   <tibble> <tibble [25 x 54]> <rowws_df [25 x 3]> <tibble [92 x 53]>
#> 2 eucalyptus <tibble> <tibble [25 x 61]> <rowws_df [25 x 3]> <tibble [119 x 53]>
#> # i 8 more variables: estimate_type <chr>, MA_mod <list>, sorensen_glm <list>,
#> #   box_cox_rating_cont <list>, box_cox_rating_cat <list>,
#> #   box_cox_rating_cat_no_int <list>, uni_mixed_effects <list>,
#> #   MA_mod_mv <list>
```

The output retains the nested tibble structure, now with one row per dataset and additional list-columns for each fitted model and visualisation. This design facilitates chaining together additional steps in the analysis, for example, by calling `make_viz()` to extract model summaries, fit statistics, forest plots and funnel plots for every model in each row.

```
meta_analysis_viz <-
  meta_analysis_results %>%
  make_viz()
```

Results for a given analysis output or visualisation are accessed by filtering on the `model_name` column, shown here for meta-analysis heterogeneity statistics:

```
# Compare heterogeneity statistics across datasets
meta_analysis_viz %>%
  filter(model_name == "MA_mod") %>%
  select(dataset, MA_fit_stats) %>%
  unnest(MA_fit_stats) %>%
  drop_na()
```

```
#> # A tibble: 2 x 6
#>   dataset    sigma2_1 sigma2_2 I2_Total I2_TeamIdentifier I2_TeamIdentifier/st~1
#>   <chr>         <dbl>    <dbl>    <dbl>             <dbl>                  <dbl>
#> 1 blue tit     0.0479  0.0217      96.7              66.5                   30.2
#> 2 eucalyptus   0.0194  0.00687     84.7              62.5                   22.2
#> # i abbreviated name: 1: `I2_TeamIdentifier/study_id`
```

This is the simplest use case: the same analytical pipeline applied to two research questions and associated many-analyst datasets.

3.6.4 Generating Analytical Subsets

Many-analyst studies can involve repeating the full analysis under varied inclusion criteria. In Chapter 2, the meta-analysis was repeated on data subsets after removing analyses rated

as low quality by peer reviewers, after excluding analyses with collinear predictors, and after removing extreme outliers. In conventional practice, sensitivity analyses of this kind are usually implemented as ad hoc copies of the primary analysis script. `ManyEcoEvo` instead treats subset generation as a first-class operation that expands the nested tibble by adding rows, one per new subset, so that the full downstream pipeline propagates through all combinations.

`generate_exclusion_subsets()` and its companion functions (`generate_rating_subsets()`, `generate_expertise_subsets()`, `generate_collinearity_subset()`) each add new rows to the nested tibble. Each new row carries correctly filtered `data` and `diversity_data` list-columns, matched to that subset's inclusion criteria. Starting from two separate many-analyst datasets, a typical call sequence produces multiple rows corresponding to each data subset:

```
meta_analysis_results <-
  ManyEcoEvo %>%
  prepare_response_variables(
    estimate_type = "Zr",
    dataset_standardise = c("blue tit", "eucalyptus")
  ) %>%
  generate_exclusion_subsets(estimate_type = "Zr") %>%
  generate_rating_subsets() %>%
  compute_MA_inputs(estimate_type = "Zr") %>%
  meta_analyse_datasets(
    outcome_variable = "Zr",
    outcome_SE = "VZr",
    filter_vars = rlang::exprs(
      exclusion_set == "complete",
      publishable_subset == "All"
    )
  ) %>%
  make_viz()
```

```
# The nested tibble now has one row per dataset × subset combination
meta_analysis_results %>%
  distinct(dataset, exclusion_set, publishable_subset)
```

```
#> # A tibble: 12 x 3
#>    dataset    exclusion_set publishable_subset
#>    <chr>      <chr>         <chr>
```

```
#>  1 blue tit   complete      All
#>  2 eucalyptus complete      All
#>  3 blue tit   partial       All
#>  4 eucalyptus partial       All
#>  5 blue tit   complete      data_flawed
#>  6 blue tit   complete      data_flawed_major
#>  7 eucalyptus complete      data_flawed
#>  8 eucalyptus complete      data_flawed_major
#>  9 blue tit   partial       data_flawed
#> 10 blue tit   partial       data_flawed_major
#> 11 eucalyptus partial       data_flawed
#> 12 eucalyptus partial       data_flawed_major
```

In the example above, the `filter_vars` argument to `meta_analyse_datasets()` — a list of `rlang::exprs()` calls — restricts the computationally expensive multivariate meta-regression to the nominated primary subset, whilst all simpler models are fitted across every row. Because the pipeline operates on the row structure rather than branching on dataset or subset identity, adding a new subset criterion requires only an additional `generate_*()` call; the rest of the pipeline is unchanged. In this way the analyses may be varied across datasets and subsets rather than forcing the user to repeatedly apply the same analysis. Lastly, resultant model fit statistics can be obtained for multiple datasets from this tibble:

```
model_fit_stats <- meta_analysis_results %>%
  filter(
    exclusion_set == "complete",
    publishable_subset == "All",
    model_name == "MA_mod"
  )

model_fit_stats %>%
  filter(exclusion_set == "complete", publishable_subset == "All") %>%
  select(model_name, mod_fit_stats) %>%
  unnest()
```

```
#> # A tibble: 2 x 9
#>   model_name   AIC   BIC  TAU2 CochransQ p_CochransQ df_error Omnibus p_Omnibus
#>   <chr>      <dbl> <dbl> <dbl>     <dbl>       <dbl>    <int>   <dbl>     <dbl>
#> 1 MA_mod      43.5  52.1     0     7442.           0      130  122.    2.93e-28
#> 2 MA_mod     128.  135.      0     2526.           0       78    2.16  1.42e- 1
```

3.7 Conclusion

`ManyEcoEvo` provides a structured workflow that maps analytical operations systematically across datasets and data subsets. The package's wrapper architecture is customised for many-analyst studies but also generalises to conventional meta-analyses, where sensitivity analyses repeat the meta-analysis across inclusion criteria or data subsets. Its contribution is largely architectural, making reproducible application of the analysis pipeline across many dataset variants a tractable, first-class operation, rather than an ad hoc after thought. Additionally, it serves as a data package for the 'Same Data, Many Analysts' study in Ecology and Evolution (Chapter 2).

4 A Framework for Questionable Research Practices in Ecological Modelling

Preface

This chapter presents work written in the style of a traditional thesis chapter, not yet submitted for publication. The chapter is preprinted at:

> **Gould, E.**, Fraser, H. S., Wintle, B. C., Rumpff, L., & Fidler, F. (2026). A Framework for Questionable Research Practices in Ecological Modelling. Applied Statistics. https://doi.org/10.32942/X2DQ0K

Abstract

1. Questionable research practices (QRPs) tend to produce apparently strong and conclusive findings that undermine the reliability and replicability of published literature. QRPs are ethically ambiguous researcher behaviours that are incentivised by academic incentive and reward structures. Recent metaresearch reveals that ecology is not immune to the 'reproducibility crisis' seen in other disciplines, due to similar rates of QRPs and a lack of transparency in published research. However, metaresearch to date focuses on hypothesis-testing research and treats data-dependent analytic decisions as inherently questionable. This is not a good fit for ecology and related fields that routinely use model-based approaches, where data-dependent decisions are often necessary and legitimate aspects of the modelling process.
2. To aid in understanding why and how frequently QRPs occur, and how severe the consequences might be, I develop a conceptual framework describing QRPs in ecological modelling, distinguishing questionable from legitimate data-dependent decisions. I present a typology of QRPs organised by decision-making mechanism and target,

reframing QRPs in modelling as practices that inflate perceived model credibility, rather than as producing false-positive statistical results.

3. I identified six QRP classes that may occur at various points in the modelling process: selective reporting, S-hacking (manipulating performance metrics), model fishing, sample curation, HARKing and overhyping. These practices threaten the reliability and reproducibility of model-based research by artificially inflating the apparent credibility of models.
4. I aim to raise awareness among modellers about different types of QRPs and how they might emerge in ecological modelling. I offer strategies to mitigate QRP risks, while preserving legitimate adaptive decision-making characteristic of ecological modelling.

Keywords: questionable research practices; ecological modelling; metaresearch; transparency; reproducibility; model credibility; researcher degrees of freedom.

4.1 Introduction

Self-report surveys of researchers' statistical practices suggest high rates of Questionable research practices (QRPs) in several different disciplines: psychology (John et al., 2012), education (Makel et al., 2021) and ecology (Fraser et al., 2018). QRPs are a set of methodological and statistical practices that can substantially influence research conclusions, and include practices like *p*-hacking, hypothesising after results are known (HARKing) and selective reporting of results (see Table 4.1). These practices fall into an "ethical grey zone" between ideal responsible research conduct and unacceptable research behaviours like fabrication and plagiarism (Butler et al., 2017, p. 94). Widespread QRPs, accompanied by a lack of transparency and openness in research reporting (Čulina et al., 2020), can leave disciplines at risk of a 'reproducibility crisis' (Fidler et al., 2017).

Current definitions and lists of QRPs are focused on hypothesis testing research (specifically, null hypothesis significance testing; NHST), characterising these practices as inflating the probability of false positive findings (see Table 3 in Nagy et al., 2025). This definition makes sense in this context, since this is the primary statistical estimand on which a finding is deemed 'publishable' in null-hypothesis significance testing. Indeed, Fraser et al.'s (2018) survey of ecology researchers asked how often they used QRPs documented in other disciplines, and many QRPs relate to *p*-values (Table 4.1). However, this NHST-centric focus creates particular challenges for ecology, because, although NHST is still popular (Fidler et

al., 2017; Stephens et al., 2007), model-based methods in ecology are increasingly common, especially within applied research contexts (Connolly et al., 2017; DeAngelis et al., 2021; García-Díaz et al., 2019). This Chapter treats model-based research broadly, encompassing both basic and applied studies that employ quantitative modelling to answer a research question. A quantitative model is defined here as any model that uses mathematical or statistical concepts to describe or forecast how a system behaves, with parameters and their relationships expressed as numerical quantities (Brudvig, 2017; Carmel et al., 2013; García-Díaz et al., 2019).

The emphasis of existing QRP definitions on Type I errors is unhelpful for model-based research because multiple sources and types of error may arise in the modelling process; there is model structural uncertainty, uncertainty in parameter estimates and predictions, and uncertainty in scenarios (Rounsevell et al., 2021; E. G. Simmonds et al., 2024). What constitutes an 'error', the source of that error, as well as the relative weighting of different errors, depend on some combination of the model's purpose (prediction, explanation, description), the type of model used (correlative or mechanistic) and the context for the model. Without a more encompassing definition of QRPs, ecological modellers may be inclined to think that concerns associated with QRPs and reproducibility are irrelevant, since many of the practices described as questionable within an NHST context do not directly relate to their work.

Building on these limitations, I argue that current QRP frameworks fail to address model-based research because modelling includes additional phases in analytic decision-making, introducing additional model outputs and researcher degrees of freedom. One principle underlying the classification of these practices as 'questionable' relates to data-driven decision making. In contrast to hypothesis testers, modellers make a series of analytic decisions drawing on both objective model performance criteria and subjective judgements (Babel et al., 2019; N. D. Bennett et al., 2013). Throughout this process, they iteratively shift between subjecting the model to analysis, validation, and testing, and refining it in response to those results (Getz et al., 2017). This process is non-linear, iterative, and generates many interim versions of the model preceding publication (Augusiak et al., 2014).

Rather than dichotomous inferences relying almost exclusively on *p*-values, model performance metrics include both qualitative and quantitative measures that incrementally build a subjective picture of model credibility (Augusiak et al., 2014; Hamilton et al., 2019). The

Table 4.1: *Examples and self-reported frequency of questionable research practices (QRPs) in hypothesis-testing research in ecology and evolutionary biology. QRPs are categorised as "cherry-picking," "p-hacking," and "methodologically flawed," indicated by the cherry, saw, and cross icons respectively. Data compiled from Fraser et al. (2018) and abbreviations defined by Makel et al. (2021). Ecology: $n = 494$; Evolution: $n = 313$.*

	Questionable Research Practice	Abbreviation	Ecology	Evolution
	Not reporting response (outcome) variable that failed to reach statistical significance.	Omitting non-significant studies or variables.	64%	64%
	Not reporting covariates that failed to reach statistical significance (e.g. p <= 0.05) or some other desired statistical threshold.	Omitting non-significant covariates.	45%	43%
	Reporting an unexpected finding as having been predicted from the start.	HARKing.	49%	54%
	Reporting a set of results as the complete set of analyses when other analyses were also conducted.	Omitting analyses.	53%	53%
	Rounding off a p-value or other quantity to meet a pre-specified threshold.	Rounding p-values.	27%	18%
	Deciding to exclude data points after first checking the impact on statistical significance.	Data Exclusion (ARKing).	24%	24%
	Collecting more data after inspecting whether the results are statistically significant.	Data peeking.	37%	51%
	Changing to another type of statistical analysis after the analysis initially chosen failed to reach statistical significance (e.g. p<= 0.05) or some other desired statistical threshold.	Analysis gaming.	54%	52%
	Not disclosing known problems in the method, analysis or data quality that potentially impact conclusions.	Hiding methodological problems.	20%	22%
	Filling in missing data points without identifying those data as simulated.	Filling in missing data.	5%	2%

publishability of a model-based study hinges on these model outputs and their collective contribution to perceived model credibility. Each output therefore represents a potential target for QRPs that aim to strategically inflate that credibility.

As such, I argue that a conceptual framework of QRPs in model-based research must account for certain kinds of data-dependent decisions, which are appropriate and justifiable aspects of the modelling process, while highlighting the primacy of guarding against data-dependent decision-making that might be questionable. Moreover, the conceptual framework should de-emphasise the risk of type I errors and false positive findings to account for other biases more relevant to how ecological models are evaluated and used.

This begs the question of whether QRPs are relevant to model-based research? If specific QRPs related to *p*-values do not apply, such as *p*-hacking, are there parallel or counterpart practices that do? What form might they take? And where in the research process would we be likely to locate them? This paper addresses each of these questions. My primary aim is to highlight the range of specific practices that are problematic in different stages of the modelling process, so as to identify the QRPs and associated decision-points relevant to model-based research. I aim to raise awareness among ecologists (and modellers within other disciplines) about the potential for QRPs throughout the modelling process. I hope to facilitate future attempts to estimate the severity and extent of QRPs and provide solutions to help mitigate questionable practices in model-based research.

4.2 Conditions for Questionable Research Practices in Ecological Modelling

Below I outline the conditions under which QRPs may arise in ecological modelling and give an overview of the modelling process and the 'objects' it produces (inputs, outputs, the model itself). This provides the foundation for exploring how QRPs might unfold in model-based research.

4.2.1 Identifying motivations for QRPs

Academic incentive structures create a fundamental tension between researchers' scientific ideals and institutional demands that systematically biases research conduct (O'Dea, Parker, et al., 2021). Institutions hire and promote based on grant acquisition and high publication

output in high-impact journals, while funding bodies reward novelty and journals preferentially publish 'surprising', significant results (Ioannidis et al., 2014; Parker, Main, et al., 2016; Tincani & Travers, 2019). As researchers in a "publish or perish" research culture wherein professional success depends on publishing, researchers are incentivised to engage in (questionable) research practices that maximise publication success over methodological rigour, transparency and reproducibility (Agnoli et al., 2017; Nosek et al., 2012) - typically without explicit awareness due to unconscious biases (Agnoli et al., 2017; Gelman & Loken, 2013; Hindman, 2015). Such practices effectively move unpublished research out of the 'file drawer' drawer' (Lortie et al., 2007; McElreath & Smaldino, 2015) and into the published literature (Ware & Munafò, 2015).

In hypothesis-testing research, most QRPs primarily affect p-values, as they determine whether the results are deemed credible and publishable. However, model-based research generates a range of different quantitative and qualitative outputs instead of, or in addition to *p*-values. It follows then, that identifying analogous components and their attributes,which determine whether the model is 'fit-for-purpose,' may reveal how and when QRPs emerge in the modelling process, and what forms they might take. These attributes become potential targets for the QRP classes I later identify in the typology (Table 4.2).

Model Advantage and Novelty

Novelty is an important factor influencing the funding and publishability of modelling research. Publication bias and funding stipulations reward and require advantage over existing approaches; either through development of new methodological approaches, enhanced performance of existing algorithms and modelling methods, or application of existing models to new contexts, such as new environmental conditions or ecological contexts (Alexandrov et al., 2011). Conversely, publication bias disincentivises the evaluation and testing of existing published models (Babel et al., 2019; Houlahan et al., 2016). This requirement for novelty incentivises model fishing and selective reporting, where researchers may opportunistically explore new modelling approaches until they achieve apparent superiority over baseline approaches.

Model Credibility

Unlike the binary nature of statistical significance, model credibility builds incrementally through multiple performance metrics and evaluation approaches (Figure 4.1, model

outputs). This multi-faceted assessment creates numerous opportunities for statistic hacking, or "S-hacking" (Table 4.2 and Table 4.3), where researchers can manipulate i) performance metric selection and thresholds, ii) validation approaches and data partitioning strategies, and iii) evaluation timeframes and spatial scales. The subjective nature of credibility assessment also fosters *overhyping* (Table 4.3), where model capabilities are overstated beyond what evaluation results justify.

Model credibility is based on the subjective degree of confidence that both the model and model-derived inferences about the real system can be used as claimed - i.e., that the model is 'fit-for-purpose' (Augusiak et al., 2014; Guisan & Zimmermann, 2000; Hamilton et al., 2019, 2022; Risbey et al., 2005). Concretely, this asks, can the model adequately answer the stated goal or research question (Christin et al., 2020), or can it be used reliably to inform management decisions (Alexandrov et al., 2011; Hamilton et al., 2019)? Credibility emerges gradually throughout the modelling process, by demonstrated adequacy (conceptual validity and predictive accuracy, Rykiel Jr., 1996) and reliability (consistent performance and transferability, Schmolke et al., 2010; Yates et al., 2018).

Fit-for-Purpose: How Purpose and Context Shape Desirable Model Properties

The purpose and context of a model do not merely describe the setting in which it is used — they actively determine which model properties are most important, and therefore what 'good performance' means in practice (N. D. Bennett et al., 2013; Eker et al., 2018; Hamilton et al., 2019). I distinguish three high-level modelling purposes following Tredennick et al. (2021): exploration, aimed at discovering meaningful patterns in data and generating hypotheses; inference, which asks specific questions about processes and relationships, evaluating the strength of evidence for statements about nature to develop system understanding; and prediction, which generates anticipatory or explanatory predictions about unobserved times, spaces, species, or other outcomes of interest.

The model's purpose strongly determines both the expected properties of model performance and the criteria used to evaluate whether the model is fit for that purpose (Bodner et al., 2020; Hamilton et al., 2019), because different purposes emphasise different attributes of model performance (N. D. Bennett et al., 2013; Bodner et al., 2020; Eker et al., 2018). The model's purpose therefore influences modelling decisions throughout the modelling cycle, including but not limited to, choice of model performance metric, modelling approach, candidate models, and model form (Aho et al., 2014; Allison et al., 2018; Babel et al., 2019;

Christin et al., 2020; Jakeman et al., 2008; Tredennick et al., 2021). For models used in basic research for inference, low bias in parameter estimates is paramount: the estimand of interest is the parameter $\hat{\theta}$ itself — a slope, an effect size, a rate — and a biased estimator produces misleading conclusions about mechanisms and relationships regardless of how well the model fits the data (Shmueli, 2010). Interpretability and mechanistic fidelity are similarly critical in this context (Yanco, 2020), because the goal is to develop and test causal explanations, and parameters must be meaningfully interpretable in light of theory.

By contrast, for predictive modelling — whether used for theory testing in basic research or for generating forecasts in applied contexts — the estimand of interest shifts to the predicted value $\hat{y}$, and the bias–variance trade-off is reconfigured accordingly (Shmueli, 2010; Tredennick et al., 2021). Some bias may be acceptable, even desirable, if it reduces variance and improves out-of-sample predictive accuracy, as in regularisation approaches. Interpretability, whilst useful, is not a necessary condition for predictive credibility — black-box algorithmic approaches may sacrifice interpretability for predictive accuracy when the goal is accurate forecasting rather than mechanistic understanding (Lucas, 2020). Generalisability and transferability become critical in predictive contexts, particularly for anticipatory predictions that must perform reliably across novel times, spaces, or species (Bodner et al., 2020; Yates et al., 2018), whereas they are of lesser concern for a model built to answer a specific inferential question in a specific system.

In applied and decision-support contexts, well-calibrated uncertainty may be more important than predictive accuracy, because decision-makers require reliable representations of uncertainty to weigh risks and trade-offs (Dietze et al., 2018a; Rounsevell et al., 2021) — a model that is accurate on average but systematically overconfident in its uncertainty bounds can be more consequential in a decision context than a less accurate but well-calibrated one. Stakeholder interpretability and acceptability also become necessary conditions for credibility in these settings, since a model that is not understood or trusted by its intended users cannot fulfil its purpose regardless of its technical performance (Badham et al., 2019; Eker et al., 2018; Hamilton et al., 2019).

4.2.2 Which Modelling Objects (Inputs, Outputs, the Model Itself) Are Affected By QRPs?

The model outputs most vulnerable to QRPs vary according to modelling context and purpose. Different modelling contexts emphasise different model objects and evaluation criteria, each of which may be manipulated to artificially inflate perceived model credibility, relative to the desired properties for the task at hand. A suite of modelling objects is generated throughout the modelling process (described below), collectively building a subjective picture of the credibility of the model. These outputs may be manipulated to improve the chance of model acceptance or publication. Models that serve different purposes are vulnerable to different QRPs and depending on the combination of modelling approach, model purpose and model type. QRPs will target different model objects (Hoffmann et al., 2021), for instance, when the modeller's estimand of interest (the "target quantity to be estimated in an analysis," Borger & Ramesh, 2025, p. 2) is a parameter estimate, as in many cases of explanatory modelling in ecology, then QRPs are likely to affect parameter estimates, parameter uncertainty, goodness-of-fit metrics, or variable importance measures. While for the purposes of prediction, QRPs are more likely to affect model components, like forecast accuracy metrics and measures of model transferability. Different questionable practices are concentrated at different locations across modelling phases (Figure 4.2).

To help conceptualise where in the modelling process QRPs might emerge, and which 'model objects' QRPs may affect, I first give an overview of the modelling process, articulating various inputs and outputs, including the model itself, model fit statistics, summary measures and other evaluation results (Figure 4.1), to which I ascribe mathematical notation.[1] This framework will also provide the foundations for extending Gelman and Loken's (2013) mathematical formalism to modelling (which I do in Section 4.4).

For simplicity, I distinguish three types of modelling (correlative, mechanistic, agnostic, following Hoffmann et al., 2021), acknowledging that there are many ways of categorising ecological models. Briefly, correlative models represent associations between two or more variables using mathematical or statistical functions fitted to data to describe or forecast underlying biological processes or system behaviour (García-Díaz et al., 2019; C. R. White & Marshall, 2019), for example, correlative Species Distribution Models/Ecological Niche

[1] Boldface notation represents a vector or a set, indicating where multiples of those objects could be generated, e.g. there may be multiple ways to operationalise a conceptual model.

Models (e.g. Briscoe et al., 2019; Sillero, 2011) and mixed-effects linear regression models (Silk et al., 2020). Mechanistic models, or 'process models,' explicitly represent components and processes of ecological systems that are theorised or hypothesised to give rise to the data (Augusiak et al., 2014; García-Díaz et al., 2019; C. R. White & Marshall, 2019; Yates et al., 2018), e.g. individual-based models or mechanistic niche models of functional traits (Kearney et al., 2019; Railsback et al., 2025). However, most models lie on a spectrum somewhere between correlative and process-based (C. R. White & Marshall, 2019). Hoffmann et al. (2021) identifies a third type, 'agnostic models,' which are non-parametric models where the functional form is not pre-specified but is instead estimated by fitting a pre-determined model class to the data (Lucas, 2020), and includes machine learning and deep learning models, as well as non-parametric correlative models like splines and Gaussian process models (Lucas, 2020; Pichler & Hartig, 2023). Agnostic models differ from both correlative and mechanistic models in that the model and its outputs are weakly theory-driven and derived from data rather than expert opinion (Lucas, 2020). In combination with the modelling purpose, each modelling type is focussed on a different 'estimand of interest' or "target quantity to be estimated" (Borger & Ramesh, 2025, p. 2).

I have divided the modelling process into three phases; 1) model construction, 2) model evaluation, and 3) model application. These distinctions align with the phases underpinning the preregistration template in Chapter 5 (Appendix D).

Conceptual Model, M_c

To begin the model development process, a conceptual model, M_c or *candidate set* of models, $\mathbf{M_c}$ is specified by the modeller, synthesising their understanding of the ecological system. Conceptual models may be represented by a set of qualitative statements, mathematical formulas, or else visually as plots or directed acyclic graphs (Shmueli, 2010). A candidate set of multiple models at this stage may represent competing hypotheses, where differences in the structure and/or parameterisation of the models represents critical uncertainty about the ecological system.

Specified Model, M_s

Next, the modeller formalises each conceptual model mathematically or statistically, M_s (Figure 4.1). The modeller chooses which variables should be included in the model, how

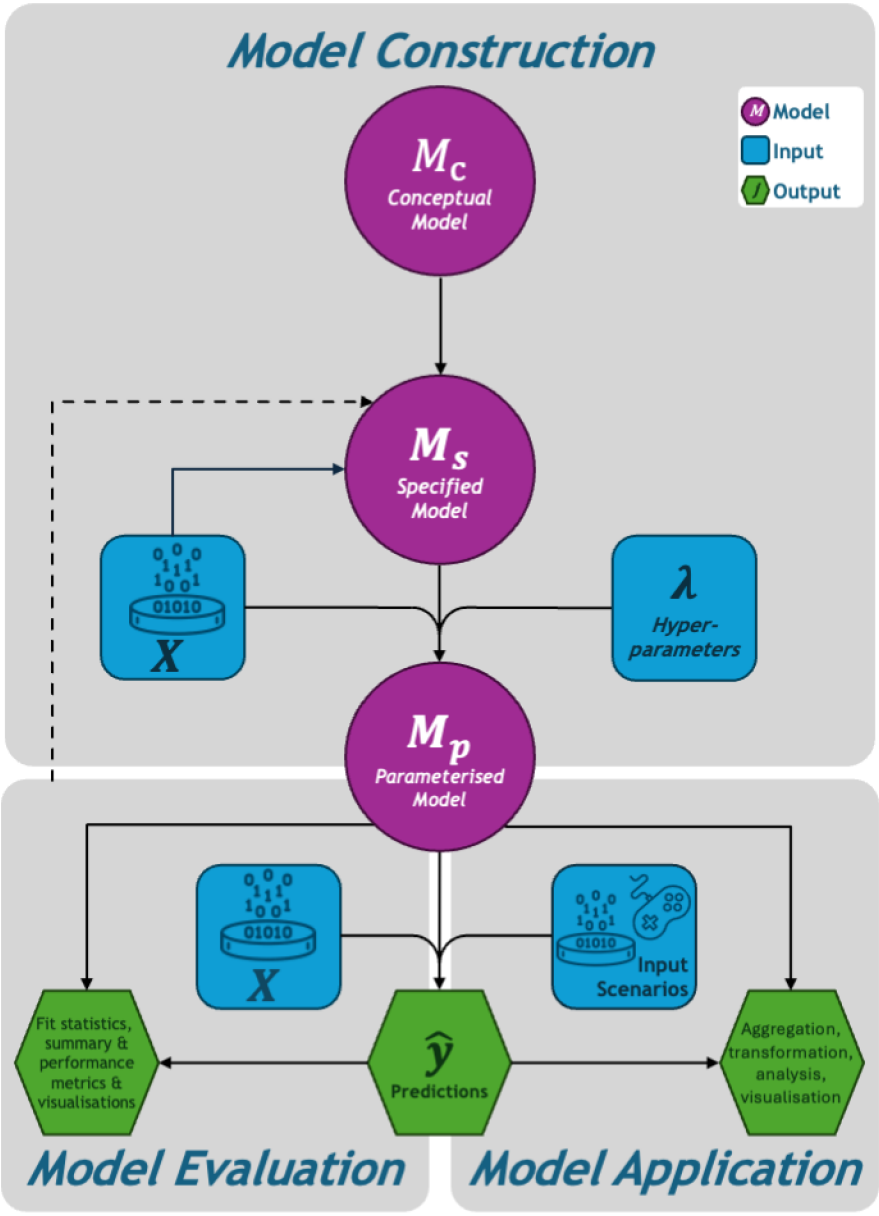


Figure 4.1: *Three phases of model development: model construction, where the conceptual model M_c is specified into the formal model ($M_c \rightarrow M_s$) then parameterised ($M_s \rightarrow M_p$); model evaluation, where the calibration and validation fits are evaluated, possibly leading to re-specification and re-parameterisation (dashed arrow); and model application, where the model is analysed to answer the research question. Modelling generates objects, including the conceptual, specified and parameterised model, ($\mathbf{M} = \{M_c, M_s, M_p\}$, purple circles); model inputs (blue squares), including hyper parameters λ and calibration settings, data $\mathbf{X}$ for model parameterisation, evaluation and application. Model outputs (M_j green hexagons) include model predictions $\hat{y}$, which are used to characterise model performance during model evaluation or subject to further aggregation, transformation, analysis and visualisation during model application. Note: Data analyses may also inform model specification during construction. New or alternative input data may be used during scenario analysis to make predictions or projections about how the system will respond to intervention $\hat{y}$.*

to operationalise or represent them in the chosen framework, what the appropriate dependencies are between variables and the model type, and the functional form of the model (if relevant). Because the variables in the conceptual model are not directly observable, they are operationalised into measurable outcomes Y and input variables X in a data matrix $\mathbf{X} = X_1, X_2, \ldots, X_p$, where p is the number of input variables, and f represents a function relating Y to X such that $E(Y) = f(\mathbf{X})$ (Hoffmann et al., 2021; Shmueli, 2010). Note that, for some predictive modelling contexts, such as data-driven modelling employing black-box algorithmic approaches, like machine-learning, f may not be specified and is instead represented by $\mathscr{I}_\lambda$ where $\mathscr{I}$ represents some learning algorithm and λ denotes its hyperparameters (following Bischl et al., 2023). Exploratory analyses are often conducted at this stage to inform variable selection, for example by analysing variable importance and examining collinearity among variables (Kass et al., 2025).

Parameterised Model, M_p

Next, each specified model M_s is parameterised yielding M_p (Figure 4.1). Model parameters refer to any component of a model that can be quantified or estimated, such as slopes or intercepts in a linear regression or growth rate in a population model (García-Díaz et al., 2019, p. 2). Regardless of the overarching model purpose, for correlative or agnostic models, parameterisation typically occurs by *estimation*, or *calibration*, whereby the modeller applies techniques, like maximum-likelihood estimation or Bayesian inference, to the data matrix $\mathbf{X}$ (Figure 4.1) to estimate the parameters specified by f, with uncertainty (García-Díaz et al., 2019; Hoffmann et al., 2021), yielding $\hat{y} = \hat{f}(\mathbf{X})$. In the case of agnostic models, the algorithm $\mathscr{I}_\lambda$ returns the fitted model and its parameters when applied to $\mathbf{X}$, $\hat{y} = \hat{f}_{\hat{\lambda}}(\mathbf{X})$. Parameters of mechanistic models are typically provided as inputs to the specified model f, gleaned from expert knowledge, published literature or via calibration (Hoffmann et al., 2021).

Irrespective of the model type, when conducting inference or explanatory modelling, the estimand(s) of interest are the parameters $\hat{\boldsymbol{\theta}}$, like standardised mean differences, correlation coefficients or response ratios (Williams et al., 2025), whereas for predictive modelling, predicted values $\hat{y}$ constitute the estimand(s) of interest (Hoffmann et al., 2021; Shmueli, 2010; Tredennick et al., 2021). However, different types of models are more likely to be used for inference or prediction in practice, for example, agnostic models are more likely to be used for prediction, although inferences about parameters are certainly possible (e.g.,

response curves / partial dependence plots from boosted regression trees, Elith et al., 2008; Lucas, 2020; Yu et al., 2021). Note that agnostic modelling approaches require the modeller to supply hyperparameters $\boldsymbol{\lambda}$ (Figure 4.1), which may be decided by the modeller, or else estimated by some tuning or optimisation method $\hat{\boldsymbol{\lambda}}$. Hyperparameters may influence the model learning process, such that with each set of hyperparameters the model will provide a different set of results (Ahmed et al., 2025).

Model Evaluation: Characterising model performance and fitness for purpose

Arriving at the preferred model or collection of final models is typically iterative, determined by the outcomes of model validation and evaluation (Shmueli, 2010). During this process, the model is subjected to a series of analyses that generate performance measures used to establish its validity, reliability, and credibility, thereby ensuring that it is fit for its intended use (N. D. Bennett et al., 2013; Eker et al., 2018; García-Díaz et al., 2019; Rykiel Jr., 1996).

I distinguish between model validation and model evaluation. Model validation checks that the fitted model M_p suitably approximates the data $\mathbf{X}$, and is evaluated using goodness-of-fit tests, and model diagnoses like residual analyses (Shmueli, 2010). Model selection narrows the candidate set of models into a single 'best' model or a smaller subset of 'best models' (e.g. AIC within $\Delta 2$), after which the modeller might choose to consider multiple models or conduct model averaging. Outside of formal model selection approaches, the procedure for determining the best model(s) may involve a degree of trial and error of different model structures that is not always preplanned (i.e. new M_s are specified after validation).

Models used for inference are assessed on how well M_p fits or explains the data $\mathbf{X}$, informed by measures like R^2, statistical significance and F statistics, whereas predictive models are additionally assessed on their predictive power or accuracy on new observations of $\mathbf{X}$ (Shmueli, 2010). Depending on the type of prediction (e.g. classification, predicted probabilities of different states or classes, ranking of probabilities or point predictions) different performance metrics will be appropriate (Shmueli, 2010).

Following validation, model evaluation of the best model(s) is undertaken, assessing the fitness of M_p for purpose by calculating additional performance measures to fully understand the model's capabilities, like constructing confusion matrices or calculating omission and commission rates. Ultimately, model evaluation is case-dependent and context-specific. The overall evaluation process, types of analyses, performance metrics, estimand of interest,

and desirable model properties all vary depending on the overarching modelling purpose, model type, and modelling approach (Bokulich, 2013; Tredennick et al., 2021).

Model Application and Analysis

Once M_p is considered plausible and fit for purpose, the modeller shifts to model application (Engelschalt et al., 2023), querying the model and using the model to undertake analyses that inform the stated research questions (Figure 4.1). Prior to analysis, model output may be subject to further processing, for example, continuous predictions may be aggregated or transformed into binary predictions for visualisation and communication purposes (Feng et al., 2019). Explanatory model output may be visualised with coefficient plots, or effect size plots to inform the relevance of observed effects (Lüdecke et al., 2020). In applied settings, forecasts or spatial projections are generated from the model based on plausible ecological scenarios or to simulate outcomes under different management actions or policies (Paniw et al., 2023), which may be subject to a range of visualisations (e.g., Barros et al., 2023, fig. 2). In decision-support settings, model application may also encompass structured decision analysis, where model outputs serve as inputs into a decision model – for example, informing rankings of management actions, resource allocation, or policy selection. In these contexts, the focus is not a predicted value per se but a recommended action or strategy, and the model outputs subject to QRPs extend accordingly.

To summarise, a collection of model outputs are generated in modelling, which may variously be the target of QRPs, including: point-estimates; such as means, medians and effect-sizes; uncertainty measures, like confidence intervals, prediction intervals, standard errors; model performance metrics, like R^2 / AIC / BIC; inference results, like p-values, credible intervals, and significance determinations; predictions and forecasts, like future values or classification outcomes; or, the model itself.

4.3 Mapping QRPs onto the Modelling Process

Here, I present a taxonomy and map of QRPs derived from the modelling literature, which aims to illustrate the different types of QRPs that might occur throughout different points in the modelling process. I follow with a synthetic example that reveals how these different types of QRPs might look in practice (Box 1).

4.3.1 Methods

To identify potential QRPs in ecological modelling and their location in the modelling process, I conducted a literature search as follows. First, I reviewed the published literature on QRPs in null hypothesis significance testing research to generate an initial list of search terms. These terms were then used in targeted searches across multiple modelling domains; including predictive modelling, habitat modelling, Species Distribution Modelling, Ecological Niche Modelling, ecological modelling, and environmental modelling. The searches used a standardised set of search string templates (e.g. `modelling area AND questionable research practice`, `modelling area AND researcher degrees of freedom`, `modelling area AND good modelling practice`, `modelling area AND bias`; see Appendix C.1 for the full list of search strings).

I inferred QRPs from practices described by authors using explicit value judgements, such as "good," "best," "poor," or "bad" practice. I included practices that were commonly or routinely conducted and where authors argued strongly for changes in research practice, and excluded practices pertaining to fraud, deliberate misconduct, or nefarious intent.

For each identified practice, I recorded: a description of the practice, the reason for its classification as questionable, any empirical or quantitative evidence of its occurrence or consequences, the phase and sub-phase of the modelling process in which it occurs (identified in Chapter 5), and the model object it targets (input, output, or the model itself, Figure 4.1). I then grouped practices consisting of the same family of research behaviours into broader QRP classes following the approach of Nagy et al. (2025), where I adopted or modified existing umbrella terms where applicable (e.g., Liu et al., 2020), and created new terms where no suitable analogue existed.

The full coded dataset is publicly archived at `doi.org/10.26188/30773906.v1`, with a human-friendly formatted version available at `doi.org/10.26188/30773831.v1`. The complete list of coded QRPs is presented in Table C.1 in Appendix C.1.

4.3.2 Results

I identified six classes of QRPs: sample curation, model fishing, selective reporting, S-hacking, overhyping, and HARKing (Table 4.1). All classes of QRPs have analogous practices

under NHST, but the practices themselves are not directly comparable. The list of QRPs I identified is not exhaustive and instead illustrates a range of practices that can occur in model-based research (See Appendix C.1, Table C.1 for the complete list).

Paradigmatic examples of QRPs are presented for each class in a roadmap (Figure 4.2), illustrating that QRPs occur throughout all phases of the modelling process, and some may occur at multiple stages. Sample curation, S-hacking and fishing were the classes of QRPs most likely to occur during model construction. The iterative nature of model validation and evaluation creates multiple opportunities for opportunistic optimisation of apparent model performance during model evaluation, with selective reporting, S-hacking and model fishing primarily affecting this phase of modelling. Fewer, but distinct, QRP types were identified for the model application phase, primarily concerning the misrepresentation of model capabilities and findings.

The target objects affected by QRPs across modelling phases reflected the focus of modelling activities and tasks, with QRPs affecting the model itself occurring primarily during model construction and evaluation, and QRPs affecting the model outputs being concentrated in model evaluation and model application phases. Below, I briefly describe the different classes of QRPs, providing illustrative examples and explaining how they can bias results.

Selective Reporting

Selective reporting involves failure to disclose methods and/or results. Selective reporting can be distinguished from other practices, such as S-hacking and model fishing, in that it lends unwarranted credibility to the model, but the model and model outputs remain unaffected. Instead of analytic decisions being data-dependent, *the communication of those results* is data-dependent. The 'garden of forking paths' is not altered by selective reporting but rather is not fully transparent.

Model Fishing

I distinguish 'model *fishing*' from the methodological technique of 'model *dredging*' for the purposes of exploration or model selection. In formal model selection procedures that employ dredging, models are selected according to a priori objective criteria. The candidate model space is typically, though not always, constrained by theoretically or ecologically motivated a priori specification.In contrast, when conducting model dredging for the purposes of exploration in pursuit of generating new hypotheses, the initial model

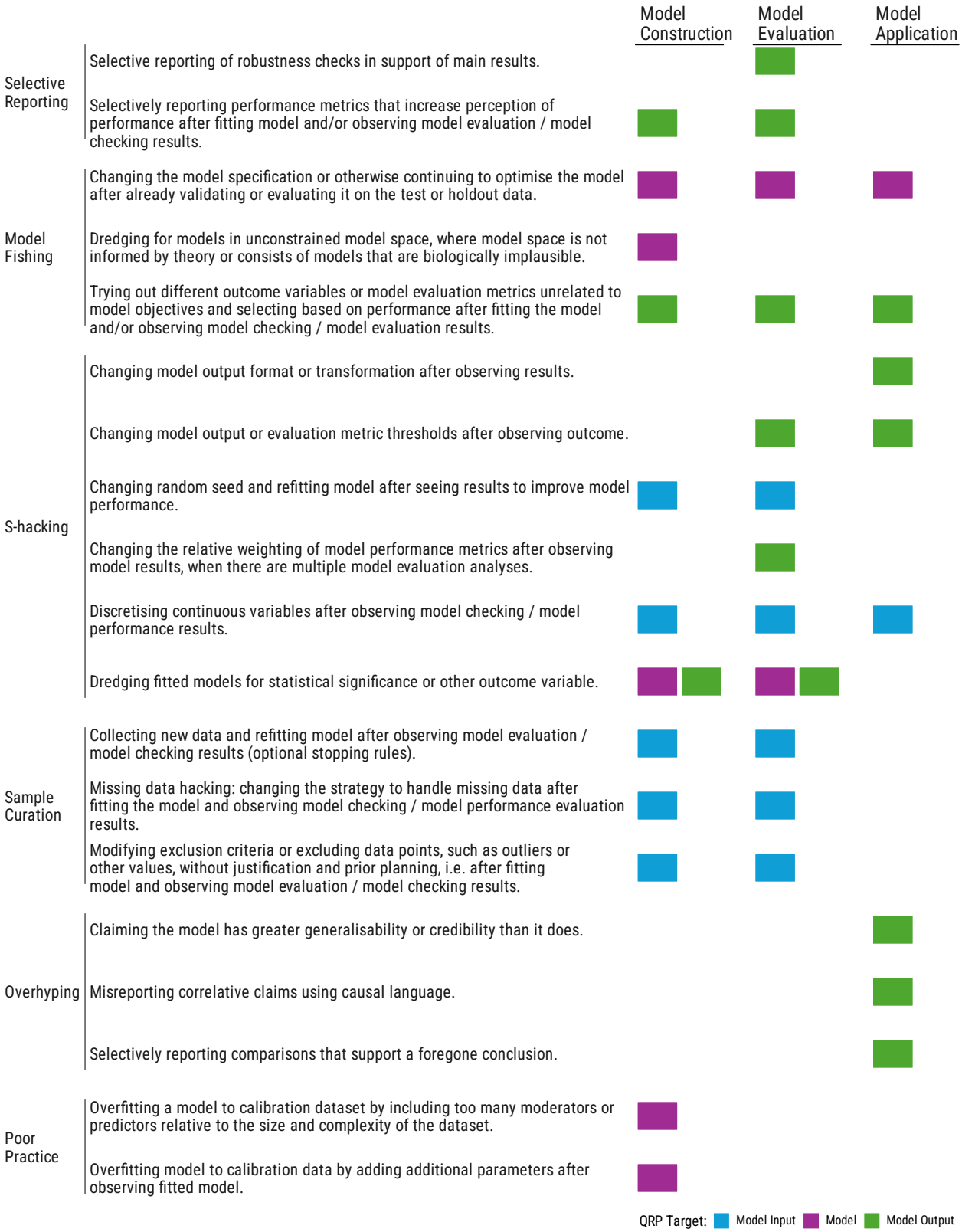


Figure 4.2: *Synthesis of questionable research practices (QRPs) in ecological modelling. QRPs may target model inputs (blue squares), the model itself (purple squares), and/or model outputs (green squares), and may occur at different phases in the modelling cycle. QRPs are grouped according to broader classes defined in Table 4.3. See Table C.1 for the full list of QRPs identified.*

space may not be as constrained, but the dredging procedure is transparently reported, and the exploratory nature of the modelling exercise is disclosed and not misrepresented post-hoc as otherwise. Model fishing occurs when the dredging procedure is not disclosed, and/or there is no formal criterion for model selection, and the overarching purpose is not exploration. Alternatively, model fishing can occur without dredging through a large model space, but by conducting alternative analyses or new model variations and selectively reporting only those with favourable results. Model fishing is problematic because of the risk of cognitive biases, such as hindsight bias, where post-hoc rationalisation combined with haphazard model selection leads to spuriously selected models. Model fishing therefore involves an element of systematic exploration of researcher degrees of freedom that is not necessarily planned, nor transparent.

S-hacking

I expanded the concept of *p*-hacking and termed it 'S-hacking', or 'statistic hacking', which encompasses analogous practices in modelling that target metrics that contribute to the publishability of a model. S-hacking involves an element of selective reporting, but a critical point of difference is that S-hacking includes the execution of alternative analyses and manipulation of data, models, or outputs to obtain a favourable result. For example, a modeller may systematically trial multiple different evaluation metrics, selectively reporting only those that present the model in a favourable light (Hildebrandt, 2018). In this instance the model remains unaffected by S-hacking. Alternatively, random seeds in model tuning can be changed after observing test set performance which can drastically alter model results (Liu et al., 2020). If S-hacking is performed during model construction or validation, or alternative model specifications are trialled after observing model performance results, the model itself is altered, and overfitted to the training data. If S-hacking is performed during model construction or validation, the model is overfitted to the training data and poorly generalises to new data. S-hacking artificially inflates model performance, resulting in spuriously selected models that may not reflect genuine ecological or predictive relationships. Any performance metric with a threshold dependent outcome (e.g. AUC, TSS, partial ROC, sensitivity, specificity, Feng et al., 2019) will be subject to the same types of practices as *p*-hacking.

Sample Curation

Sample curation (*sensu* Nagy et al., 2025) includes a range of data-dependent decisions about model inputs without justification or prior planning, i.e. after model fitting or observing model evaluation or application results. Sample curation may include removing observations in order to make a correlation of interest become significant and generating a data-dependent criterion for the exclusion of particular observations (Nagy et al., 2025). Opportunistic handling of missing data could occur in a number of ways, for instance when a researcher attempts list-wise deletion, multiple imputation or inverse probability weighting. The expected results may only appear with one of those options, which is problematic if the researcher only reports this strategy in the paper, and omits the results from the other data handling methods (Nagy et al., 2025). Similarly, opportunistic stopping occurs when new data is collected and is used to re-parameterise the model after previously observing model validation and model evaluation results, without reporting results of earlier iterations (Table 4.3, Table C.1).

Hypothesising After Results Are Known (HARKing)

Although the problem of HARKing might seem exclusively like a problem for explicit hypothesis-testing contexts, it is important to acknowledge that ecological models implicitly encapsulate hypotheses in the form of assumptions about which patterns, relationships, or predictors are most relevant to the system being modelled (Bodner et al., 2020; Prosperi et al., 2019; Schuwirth et al., 2019). For example, the choice of which variables to include or exclude from a model are based on implicit hypotheses about which processes are relevant to the system. In the case of modelling, HARKing can occur when a researcher presents a post-hoc explanation and justification for the variables or model structure that performed best, while failing to disclose the initial exploration of other variables or model structures. As such, HARKing in ecological modelling for purposes other than hypothesis testing is likely to occur as an effect of other related QRPs (Table 4.3) rather than as the motivating practice.

Overhyping

Overhyping involves claims about the models' performance that are not substantiated by model evaluation results, such as claiming the model has greater generalisability than it does (Corneille et al., 2023). A specific form of overhyping involves misreporting correlative

claims using causal language, which is particularly common in studies evaluating conservation interventions using observational study designs (Josefsson et al., 2020). The practice of implying causation from correlation can cause false confidence in the intervention's effectiveness while ignoring the real mechanisms for the observed effect.

Box 1: Synthetic Example Of Questionable Research Practices In Applied Ecological Modelling

A modeller seeks to predict species abundance on the basis of habitat quality to help inform conservation management (Figure 4.3). When the modeller evaluates how two different management actions affecting habitat quality influence species abundance, the initial a priori model does not provide sufficient certainty for choosing between one action and another (Panel A, Stage 1). The modeller revises the model without theoretical justification, instead opportunistically trialling different models and selecting the one with the best Performance Score (Panel B). On checking the predicted species abundance for the two actions on the overfitted model, the modeller finds that the actions are still not clearly distinguishable in terms of their predicted outcomes (Panel A, Stage 2), so the modeller adjusts the scenario input values for the two management actions, and plots the predicted outcomes (Panel B, Stage 3). They are sufficiently happy that the model now clearly supports their preferred management action B and proceed to publish the overfitted model, its predictions and management recommendations without ever disclosing their model fishing and scenario hacking, effectively a form of HARKing ('hypothesising after results are known'). The impacts of the modeller's actions are summarised in Panel C.

Extended Caption for Figure 4.3:

A synthetic illustration of model fishing & scenario hacking (HARKing).

A. Violin boxplots of predicted species abundance for two management actions from two models constructed at different stages of the modelling process: a priori model (stage 1), a model generated from a model fishing exercise (stage 2), and the same model, but illustrating scenario hacking (stage 3). Dots are predicted values. Violin outlines illustrate kernel density probability distributions, where the width of the shaded area represents the corresponding proportion of data. The model-estimated median and quartiles are displayed for each action. Colours correspond to the scenario actions displayed in Panel C.

B. Performance Scores calculated from multiple model performance measures for the Initial Model and a new, superior Overfitted Model derived from model fishing (greater overall Performance Score). See Lüdecke et al. (2020) for metric calculation details.

C. Predicted species abundance as a function of habitat quality for the Initial Model (yellow line) and the Overfitted Model generated from the model fishing exercise (dashed orange line). The management action scenarios used in the first two stages of modelling are shown as solid light blue and light green lines. Scenario hacking occurs when the modeller selects two new management scenarios with a greater difference in mean predicted species abundance under the Overfitted Model. See Appendix C.2 for code.

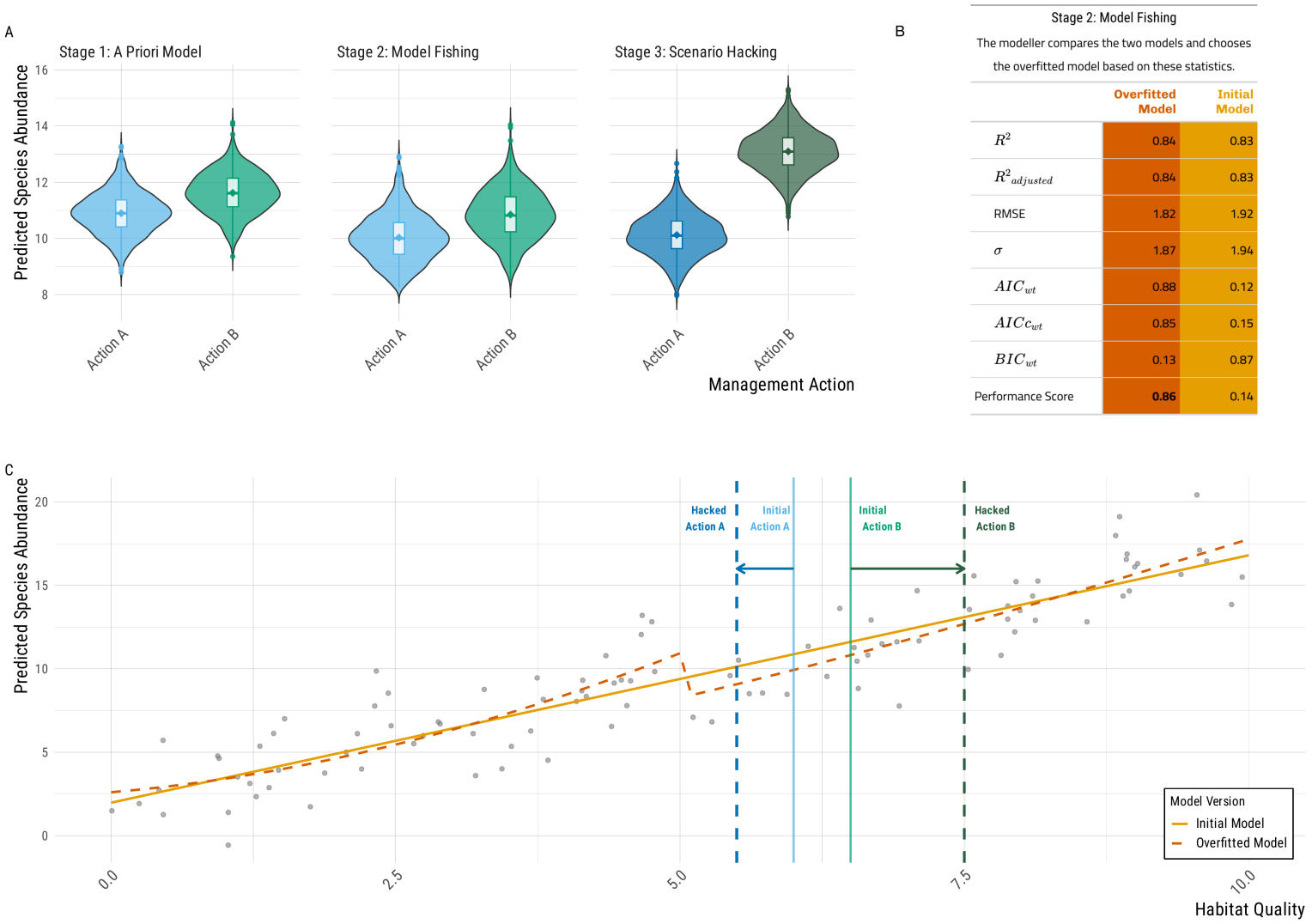


Figure 4.3: *Illustration of model fishing and scenario hacking (HARKing) with a synthetic example. Extended caption is located in Box 1.*

4.4 Formalising the "Garden of Forking Paths" in Model-Based Research

In this section, I extend Gelman and Loken's (2013) mathematical formalism explaining the emergence of QRPs, or the "garden of forking paths" (introduced in Chapter 1) to model-based research. Outlining the mathematical formulation of QRPs for ecological modelling helps to formally differentiate defensible and questionable data-dependent decisions.

As emphasised above, I hesitate to designate all data-dependent analytic decisions as *questionable*, as is implied in the prevailing literature on preregistration and QRPs. There are situations in modelling where decisions are necessarily dependent on the outcome of previous analytic decisions within the modelling workflow (Liu et al., 2020), and so not all data-dependent analytic decisions are automatically *questionable* within the context of modelling. For example, many modelling decisions are data-driven, like the choice to remove correlated variables or checking for distributional assumptions to aid in deciding the most appropriate model functional form (See Chapter 5, Figure 5.4 for an example from the case study). Liu et al. (2020) distinguish *defensible* from *questionable* motivations for engaging in data-dependent analytic decision-making, by classifying them as either *systematic* or *opportunistic*, respectively.

Box 2 helps us to *formally* distinguish between defensible and opportunistic data-dependent decisions, for both analytic and reporting decisions. This, in turn, helps to identify and distinguish between different types of QRPs in model-based research.

Box 2: Formal Description of Questionable Research Practices

The modeller makes a series of analytic decisions to derive M_p from M_s, from M_c, referred to hereafter as M for simplicity (see Section 4.2.2 for notation definitions). I term the sequence of modelling choices throughout the modelling process the realised "modelling path." Analytic uncertainty, or analogously 'researcher degrees of freedom,' propagates combinatorially along each decision-point to inform a multiplicity of plausible analysis strategies (Hoffmann et al., 2021) constituting the "garden of forking paths" (Gelman & Loken, 2013). Consider that a modeller faces some decision C along that path about a modelling task concerning model M and some observed data $\mathbf{X}$, with a predetermined choice or decision rule ϕ. Decisions made before observing

data or model outputs reflect idealised practice where choices are predetermined and independent of results $C(\mathbf{X}, M; \phi)$. When the modelling choice is "data-contingent" $\phi(\mathbf{X}, M, M_j)$ insofar as it hinges on the observed state of the model(s) M and/or any associated outputs M_j and data $\mathbf{X}$ at that point along the modelling path, it satisfies a broad definition of 'questionable.'

I define defensible data-dependent decisions ϕ_D as following a systematic process $\Psi_{\text{systematic}}$:

$$\phi_D(\mathbf{X}, M; \Omega, \Psi_{\text{systematic}}) = g(\mathbf{X}, M, \Omega) \tag{4.1}$$

Where $g(x)$ is a deterministic function of the data, model state, and predefined decision-rule Ω, based on systematic objectives such as, model adequacy, predictive accuracy or theoretical consistency.

In contrast, I define questionable practices ϕ_Q as:

$$\begin{aligned} \phi_Q(\mathbf{X}, M; \Psi_{\text{opportunistic}}) = \phi^* \\ \text{such that } h(\phi^* | \mathbf{X}, M, R^*) \geq h(\phi | \mathbf{X}, M, R^*) \\ \text{for all } \phi \in \Phi \end{aligned} \tag{4.2}$$

Where the decision-making is opportunistic and result-seeking $\Psi_{\text{opportunistic}}$, and $h(\phi | \mathbf{X}, M, R^*)$ represents how well decision ϕ serves the desired outcome R^*. Data-contingent decisions are therefore *questionable* when undisclosed desired outcomes R^* - whether driven by publication incentives, stakeholder expectations, or other motivating pressures – influence the direction that the realised modelling path takes.

A defining aspect of QRPs is that they remain undisclosed. Just as decisions about the modelling process can be questionable, so too can reporting practices. I apply the same logic describing questionable modelling practices to reporting practices:

$$C_{\text{Reported}} = S(C(\mathbf{X}, M, \phi)) \tag{4.3}$$

Where S is a selecting function that determines what to report from a set of conducted analyses.

Journal method or article length conventions restrict complete transparency, and not all results can be reported. The decision about what to report from a set of conducted analyses S is made following predetermined plans Ω, not contingent on observed results S_{pre}:

$$S_{\text{pre}}(C(\mathbf{X}, M, \phi)) = s(C(\mathbf{X}, M, \phi), \Omega) \tag{4.4}$$

Questionable reporting S_Q, or as it is more commonly known, *selective reporting*, occurs when the reporting is opportunistic and contingent on the observed results, and optimised for desired outcomes R^*:

$$\begin{aligned} S_Q(C(\mathbf{X}, M, \phi); \Psi_{\text{opportunistic}}) = C^* \\ \text{such that } h(C^*|R^*) \geq h(C|R^*) \\ \text{for all } C \in C(\mathbf{X}, M, \phi) \end{aligned} \tag{4.5}$$

I avoid defining $\Psi_{\text{systematic}}$ reporting decisions, as I did for modelling decisions, and instead advocate modellers prespecify what results will be reported.

4.5 A Typology of QRPs

Here, I present a typology of QRPs from which I designate the practices as questionable or defensible (Table 4.2). The typology considers combinations of the decision-making mechanism (a priori, defensible and questionable data-dependent decision-making), the target T of the practice (the model M or model outputs M_j), and the nature of reporting (prespecified or selective). This allows us to account for QRPs where the model and/or model outputs are directly affected by the questionable practice such that their realisations are different from what would have been observed if the practice was not undertaken (i.e. M^*, M_j^*), as well as QRPs where the model and outputs remain unaffected, but are selectively reported.

Table 4.2 summarises the key distinctions resulting from the workings in Box 2, helping to distinguish between questionable and defensible data-dependent decision-making, and to identify different classes of QRPs, with which I provide formal examples. The rows in Table 4.2 discriminate between a priori and (defensible versus questionable) data-dependent decisions, while the columns distinguish between prespecified and selective reporting.

No questionable practices occur when analytic and reporting decisions are made a priori (first row, first column) representing an idealised scenario (e.g. preregistered analyses), which is difficult to implement in practice for most ecological modelling. Moving across to the right, all decisions about the modelling and analysis are made a priori, but the results are selectively reported to improve the apparent suitability of the model to the analysis problem (questionable).

The subsequent row represents situations where there are data-dependent choices made by the modeller, representing most situations in ecological modelling. In the first instance, there is some process stipulated a priori for deciding on how the modeller will resolve any data-dependent decisions (e.g. 'registered flexibility,' Chapter 5, Section 5.2.3), and it is already decided what results will be reported (defensible). Moving to the next column over, the modeller uses registered flexibility to inform modelling choices, but in this case, they selectively report some results (questionable). In the first two rows, choices about the model and modelling analysis remain unaffected by QRPs, even when selective reporting occurs.

The final row of Table 4.2 indicates QRPs where data-contingent decisions are optimised for preferred results R^*, affecting either the model and/or modelling outputs. Because QRPs involve a degree of non-disclosure and intransparency by definition (Definition 4.5.1), I have merged the two columns that distinguish between the presence of selective reporting. In Box 1 I formally illustrate this with two example QRPs, model fishing (affecting the model M itself) and scenario hacking (affecting the model outputs M_j only).

Definition 4.5.1. QRPs occur when a researcher makes **opportunistic** data-dependent *analytic* and/or *reporting* decisions; i.e., decisions that depend on an **undisclosed** desired outcome rather than a prespecified objective decision criterion, and which artificially inflate the apparent, accuracy, precision or performance of a model and/or its outputs.

Table 4.2: *Distinguishing between questionable and defensible motivations for decision making during modelling, and the target of the decisions. I designate questionable practices in grey shaded cells and assign practices to classes of QRPs (described in Table 4.3). See Box 2 for notation and expanded definitions.*

Analytic Decision Type and Motivation	**With Pre-specified Reporting S_{pre}**	**With Selective Reporting S_Q**
A priori analytic decision, where: $\phi(T) \in \Phi_{\text{fixed}}$	$S_{\text{pre}}(\{C(X, M, \phi\})$ $S_{\text{pre}}(\{C(X, M, \phi(M_j))\})$ Decisions affecting the model ϕ or model outputs $\phi(M_j)$, and which analyses and results will be reported, are decided a priori.	Selective Reporting: $S_Q\{C(X, M, \phi\})$ $S_Q\{C(X, M, \phi(M_j))\})$ A priori analysis decisions about the model ϕ and model outputs $\phi(M_j)$, but reporting choices are data-dependent, and are selectively reported S_Q.
Defensible analytic decision, where: $\phi_D(T) \in \Phi(\Omega)$ *Systematic evaluation according to predetermined criteria.*	$S_{\text{pre}}(\{C(X, M; \phi_D\})$ $S_{\text{pre}}(\{C(X, M; \phi_D(M_j))\})$ Data-dependent and defensible modelling decisions $C(\mathbf{X}, T; \phi_D)$, with pre-specified reporting choices S_{pre}.	Selective Reporting: $S_Q(\{C(X, M; \phi_D\})$ $S_Q(\{C(X, M; \phi_D(M_j))\})$ Modelling decisions are data-dependent and defensible, but subject to selective reporting.
Questionable analytic decision, where: $\phi_Q(T) \in \Phi(R^*)$ *Optimising for preferred results R^*, rather than assessment.*	Model Fishing: $C(X, M; \phi_Q(M))$ S-hacking: $C(X, M; \phi_Q(M_j))$ Modelling and analysis decisions are opportunistic and therefore questionable.	By definition, desired outcomes R^* and ϕ_Q are undisclosed, and may be tacit rather than explicit.

In Table 4.3, I present formal descriptions and practical examples of the different types of QRP in the typology. While some example scenarios may appear to be defensible data-dependent decisions (for example, testing different functional forms under 'model fishing'), I include corresponding formal descriptions to remind the reader of the distinction between systematic and opportunistic analytic decision-making which denotes when data-contingent decisions are questionable, or not. I elaborate on these practices in the QRP map above (Section 4.3).

In summary, based on the mathematical formalism and typology, a clearer definition of QRPs in model-based research is apparent.

Table 4.3: *Formal descriptions of QRP classes, their definitions and some practical examples.*

QRP Class and Definition	Example Scenarios	Description	Formal Description
Selective Reporting Choosing to report only certain analyses, models, metrics, model results or comparisons that yield favourable results or desired conclusions without disclosing the full range of analyses performed.			
Multiple model testing with selective reporting	• Reporting only the model with best R^2 (testing linear, polynomial, and exponential models) • Highlighting only the ecological model supporting preferred hypothesis	Testing multiple model structures $(M_1, M_2, \ldots, M_k)$ and reporting only the model that produces results most align* with desired outcomes R^*.	$S_Q(\{C(\boldsymbol{X}, M_k, \phi^{(M)})\}_{k=1}^{K}; R^*) = C(\boldsymbol{X}, M_{k^*}, \phi^{(M)})$ such that $h(C(\boldsymbol{X}, M_{k^*}, \phi^{(M)})\|R^*) \geq h(C(\boldsymbol{X}, M_k, \phi^{(M)})\|R^*)$ for all $k \in \{1, \ldots, K\}$
Multiple output evaluation with selective reporting	• Reporting only significant goodness-of-fit metrics • Highlighting only time periods showing desired trends • Emphasising model predictions supporting preferred conclusions	Evaluating multiple outputs $(M_{j_1}, M_{j_2}, \ldots, M_{j_j})$ from same model (different metrics, time periods, spatial scales, etc.) and reporting only those that align with desired outcomes R^*.	$S_Q(\{C(\boldsymbol{X}, M, \phi_j^{(M_j)})\}_{j=1}^{J}; R^*) = C(\boldsymbol{X}, M, \phi_{j^*}^{(M_j)})$ such that $h(C(\boldsymbol{X}, M, \phi_{j^*}^{(M_j)})\|R^*) \geq h(C(\boldsymbol{X}, M, \phi_j^{(M_j)})\|R^*)$ for all $j \in \{1, \ldots, J\}$
Model Fishing			
Exploring model specifications or (fitting) variable combinations without theoretical justification, seeking favourable results, undisclosed.	• Adopting the ecological process model that supports preferred management conclusions • Choosing model covariates giving desired effect direction • Settling on model complexity producing significant results rather than optimal predictive performance	Testing multiple model $(M_1, M_2, \ldots M_k)$ structures or specifications and selecting model based on results most align* with desired outcomes R^*, rather than following pre-specified model selection criteria or theoretical justification.	$\phi_Q^{(M)}(\boldsymbol{X}, M; R^*) = M_{k^*}$ such that $h(C(\boldsymbol{X}, M_{k^*}, \phi)\|R^*) \geq h(C(\boldsymbol{X}, M_k, \phi)\|R^*)$ for all k
S-hacking Statistic-hacking			
Manipulating model inputs, outputs or the model itself (random seeds, outcome variable and/or performance thresholds or metrics) to obtain a favourable value of a performance measure (statistic or metric), without disclosing.	• Reporting only best performing goodness-of-fit metrics (after testing R^2, AIC, RMSE, etc) • Selecting validation approaches (cross-validation, holdout, bootstrap) based on most accurate • Adjusting thresholds until desired statistical outcomes are achieved	Trialling different performance metrics ($M_{j_1}, M_{j_2}, \ldots, M_{j_j}$), evaluation thresholds, or validation approaches and selecting metric with best model performance R^* (not pre-specified or theoretically justified).	$\phi_Q^{(M_j)}(\boldsymbol{X}, M; R^*) = M_{j^*}$ such that $h(M_{j^*}(C(\boldsymbol{X}, M, \phi))\|R^*) \geq h(M_j(C(\boldsymbol{X}, M, \phi))\|R^*)$ for all j
Sample Curation			
Selectively including, excluding, or modifying data points in the sample used to develop or evaluate a model without disclosure, often to improve model performance (a specific instance of S-hacking).	• Removing "inconvenient" sites that don't fit expected patterns • Adjusting temporal boundaries to exclude unfavourable periods	The modeller trials different curated versions of the original dataset $\boldsymbol{X}$ in order to achieve a model that meets desired outcomes R^*, without a priori exclusion rules or statistical/ theoretical basis.	$\phi_Q^{(\boldsymbol{X})}(\boldsymbol{X}, M; R^*) = \boldsymbol{X}^*$ such that $h(C(\boldsymbol{X}^*, M, \phi)\|R^*) \geq h(C(\boldsymbol{X}', M, \phi)\|R^*)$ for all $\boldsymbol{X}' \in \mathcal{X}(\boldsymbol{X})$
HARKing Hypothesising After Results are Known			
Selected model(s) and/or results presented as if pre-specified and theoretically justified before data exploration or parameterisation, and model selection procedure is not disclosed or adequately described. May accompany fishing expeditions, model dredging analyses and model selection procedures that have been selectively reported.	• Post-hoc ecological theories used to justify variable inclusion, when exploratory analyses reveals unexpected predictive variables • Theoretical explanation created post-hoc for model structures that unexpectedly perform well against other models	Post-hoc conceptual models are constructed to explain model results, serving desired outcomes R^*.	$H_Q(\boldsymbol{X}, M, C; R^*) = H^*$ such that $\text{consistency}(H^*, C(\boldsymbol{X}, M, \phi)) \geq \text{consistency}(H', C(\boldsymbol{X}, M, \phi))$ for all $H' \in \mathcal{H}$, where: M_c^* represents the post-hoc conceptual model, M_c is the space of possible conceptual models, **consistency()** measures how well a conceptual model explains the observed results
Overhyping			
Exaggerating the capabilities, generalisability, or reliability of a model beyond what is justified by the evidence.	• Typically, overhyping features in the discussion section of a paper	—	—

4.6 Discussion

Researcher degrees of freedom threaten the credibility and reliability of model-based research, just as they do in hypothesis testing research. The findings of this Chapter underscore that researcher degrees of freedom abound in the modelling process, providing ample opportunity for QRPs that accompany researchers' drive to publish. This aligns with Liu et al.'s (2020) qualitative analysis of how researchers make analytic decisions when faced with arbitrary choices or analytic uncertainty in the context of a research culture that promotes publication bias. I showed that QRPs can occur at any point in the modelling process, and may affect different model objects, including the model inputs, the model itself, or model outputs. My analysis identified that some QRPs are analogous to those in hypothesis-testing research. However, I also demonstrated that the unique methodological characteristics of ecological modelling require QRPs to be defined on our own terms. Specifically, any such definition must account for the iterative and adaptive nature of the modelling process and the necessity of making data-contingent decisions throughout it. These particularities of modelelling practice have resulted in significant resistance to the idea of QRPs in model-based research, and to the applicability of preregistration for mitigating them in ecological modelling (Dwork et al., 2015; MacEachern & Van Zandt, 2019). I explore these tensions below, emphasising how the conceptual framework provides a way forward through the tricky problem of delineating where the concept of QRPs apply in ecological modelling, and where it does not.

4.6.1 *Transparency* Determines When 'Poor Practices' are 'Questionable'

Many of the practices I identified as 'questionable' could simply be considered 'poor practice,' especially when those practices result in biased or overfitted models. The modelling context, including constraints on feasibility, data availability and coverage, together with the model purpose (e.g. prediction versus inference) will delineate when such practices are questionable or methodologically flawed. The fundamental issue with QRPs is that they remain *undisclosed*. Given that QRPs are practices that artificially improve the way models are perceived, full transparency allows the reader to appraise the appropriateness of practices like altering data, changing model specifications, or calculating additional performance metrics, contingent on the modelling context (Woo et al., 2017). Transparent

reporting is essential for properly evaluating the credibility and suitability of the model for its intended application.

4.6.2 *Opportunism* Determines When Data-dependent Analytic Decisions are 'Questionable'

Gelman and Loken's (2013) formalism describing the garden of forking paths implies that data-dependent decisions, at least in the context of null hypothesis significance testing, are inherently questionable. This has limited relevance in ecological modelling because it is inherently adaptive. My extension of the formalism to model-based research circumvents this incompatibility by distinguishing *opportunistic* from *systematic* data-dependent decisions. Based on my formalism, I argue that data-driven analysis decisions are not inherently questionable. *Questionable* practices occur when the decision is contingent on the observed results, *and* the choice is based on how well it serves undisclosed desired outcomes, whereas *defensible* data-contingent decisions follow a prespecified decision rule.

4.6.3 Data Constraints Amplify the Risk of QRPs

The nature of ecological data confers specific vulnerabilities to QRPs during the modelling process. Small datasets are prevalent in ecology and frequently exhibit inconsistent structure, arising from intermittent or one-off data collection and spatial constraints on sampling (Todman et al., 2023). Additionally, data collection is highly constrained by budget and logistical feasibility, consequently field ecologists often take a 'kitchen sink' approach to data collection, whereby they "often measure almost everything they can" (Mac Nally, 2000, p. 669). Models that analyse small datasets are more likely to be overfitted due to the high number of parameters compared to the degrees of freedom in the data (Todman et al., 2023).

These conditions provide substantial opportunity for unconstrained dredging of model space whereby modellers include covariates with little or no theoretical justification or ecological relevance, leading to biologically implausible models being considered (Fourcade et al., 2018; Franks et al., 2025; Shmueli, 2010). Although I classified these "causal salad" approaches to modelling (McElreath, 2020) as 'poor practice' instead of QRPs, when modellers are engaged in model dredging without predetermined selection rules and criteria, the risk of both poor and questionable practices, like model fishing, is heightened under these conditions.

The same data constraints that facilitate model dredging also inhibit the detection of resulting problems. When datasets are small or incomplete, there is often insufficient data to perform model evaluation on independent data (Bodner et al., 2020; Dietze et al., 2018b; Wood et al., 2020). When models are evaluated on training data only, data leakage causes biased estimates of performance - making overfitting hard to detect (Christin et al., 2020; Kapoor & Narayanan, 2023; Lewis et al., 2023; Stock et al., 2023).

4.6.4 Impacts of QRPs

Many of the questionable practices I identified specifically target perceptions of model reliability, accuracy, and generalisability rather than traditional statistical thresholds (e.g., overhyping claims and misreporting correlative findings with causal language). This supports a model-centric definition of QRPs as practices that "artificially inflate the apparent accuracy or precision of a model, its predictions, and/or evaluation tests." The mathematical formalism in the typology demonstrates the diverse ways researchers can manipulate both technical model properties and broader perceptions of model fitness for purpose, providing a comprehensive framework for understanding questionable practices in model-based research.

QRPs collectively undermine the reliability and reproducibility of ecological modelling research in several ways:

1. **Inflated performance estimates** that do not reflect true performance and result in overfitting.
2. **Spurious model selection** that identifies models based on chance rather than plausible biological mechanisms or predictive relationships.
3. **Reduced reproducibility** due to undisclosed researcher degrees of freedom.
4. **Compromised generalisability** from overfitted models that fail to transfer to new contexts or make accurate forecasts.
5. **False confidence** in ecological understanding and management recommendations.

Overfitted models are fitted to both regular and irregular features of the sampled data but are unable to distinguish between them (Pu et al., 2019), generating spurious predictions that poorly generalise to new data (Lewis et al., 2023; Todman et al., 2023).

In applied settings, where modelling is often focused on generating anticipatory predictions to inform management or policy decisions, this is particularly problematic. Inaccurate predictions or inappropriate inferences may lead to irreversible or detrimental ecological consequences (García-Díaz et al., 2019; Wood et al., 2020). At the research–practice interface, pressures from decision-makers and funders who favour definitive conclusions over uncertain or exploratory results may function analogously to publication bias, creating particular risks of S-hacking, or overhyping. For example, non-specialist decision-makers may demand "validation" as a binary criterion, pressuring researchers to force models meet performance thresholds (Augusiak et al., 2014). Or modellers may be rewarded for findings or recommendations that align with specific interests (Jakeman et al., 2024), or appear more certain than they are to strengthen the apparent policy relevance (Josefsson et al., 2020). Practical constraints common in applied settings - like compressed project timelines and limited resources for sensitivity analysis and uncertainty quantification (Schmolke et al., 2010) — do not necessarily result in questionable research practices. However, they may nonetheless incentivise ad hoc analytic decisions over idealised a priori specification.

The prevalence of these practices suggests systemic issues in training, incentives, and quality control within the ecological modelling community. The concentration of QRPs in the model construction and evaluation phases indicates particular vulnerabilities in how models are specified, fitted, and evaluated. The large number of decision-points where researchers can exercise degrees of freedom suggest that safeguards should target these critical phases of the modelling process. It is worth noting however, that insufficient training or modelling literacy, combined with growing expectations that researchers build and apply increasingly complex models and methods, compounds the risk of QRPs incentivised by publication norms and other external pressures. Addressing modelling literacy and addressing incentive structures are therefore complementary, rather than competing, imperatives for improving the reliability of ecological modelling.

4.6.5 Potential Solutions

Raising Awareness

Awareness of the distinction between systematic and opportunistic data-dependent decisions is limited, and because some data-dependent decisions are a legitimate aspect of the modelling process, it may seem that all data-dependent decisions are acceptable. The

impact of data-dependent decisions in machine learning is increasingly understood and is encapsulated within the term 'data leakage,' whereas in ecological modelling more broadly, the equivalent problem of 'model selection bias' remains mostly overlooked (Campbell, 2021) and underappreciated within applied research contexts (Arnqvist, 2020). Here, I emphasise an intersecting problem that has been attributed as a major cause of science's reproducibility crisis, where data-contingent decisions may be opportunistically exploited to increase the likelihood of publication. I have formalised the distinction between defensible or questionable research practices, facilitating a modelling-appropriate conceptualisation of QRPs. As a first step in addressing the threat of QRPs to the credibility of model-based research, I wish to draw attention to this distinction – and the possible consequences of QRPs – among the ecological modelling community. However, given that cognitive biases are rarely deliberate, awareness alone is insufficient for preventing QRPs within a publish-or-perish research culture (Zvereva & Kozlov, 2021).

Increasing Transparency

Modelling is not typically transparent, leaving readers unable to assess whether appropriate models were used or to identify the primary research contribution (Arnqvist, 2020). Given that the threat of QRPs largely stems from a lack of disclosure, ecological modelling is at significant risk of QRPs. I echo broader calls for improving transparency in ecology (O'Dea, Parker, et al., 2021; Parker, Main, et al., 2016; Powers & Hampton, 2018), emphasising that transparency is a fundamental requirement for reducing the risk of QRPs in ecological modelling. It is acknowledged that modelling's lack of transparency is, in large part, driven by sociocultural and institutional norms that restrict the length of a paper, require a neat and coherent narrative and favour some data analysis techniques and results over others (Rijnhart et al., 2021). Broader methodological reform in research culture, as well as specific tools, are needed to achieve improvements in transparency.

Reporting checklists and guidelines outline a minimum set of methodological elements and results to include in published research and are increasingly being adopted by journals in ecology ("A checklist for our community," 2018; Fidler et al., 2018; Haddaway et al., 2018; Hillebrand & Gurevitch, 2013; O'Dea, Lagisz, et al., 2021). However, there are only several reporting checklists developed for ecological modelling (see Chapter 1), and no ecology journals have encouraged or mandated modelling-specific checklists at the time of writing. I leave the work of defining the content of reporting checklists up to the ecological modelling

community, but reiterate repeated calls in the modelling community to articulate the model's purpose, context and performance criteria, ideally before modelling begins (N. D. Bennett et al., 2013; Jakeman et al., 2006; Wood et al., 2020). This chapter illustrates when practices are questionable, and that many QRPs target or alter model performance metrics – either through direct manipulation of the model and model outputs, or through selective reporting. Specifying these decisions a priori and reporting them reduces inadvertent engagement in QRPs and equips readers to evaluate the risk of QRPs.

Decision-Support Modelling as a Model for Good Practice

Whilst decision-support modelling introduces additional opportunities for QRPs — as discussed above and flagged for future investigation in Section 4.6.6 — the norms developed within this context offer instructive lessons for ecological modelling more broadly.

The first transferable lesson is explicit purpose specification: defining, before modelling begins, what question the model is intended to answer and what properties it must therefore possess to be adequate (N. D. Bennett et al., 2013; Jakeman et al., 2006; Schuwirth et al., 2019). Without clearly specifying the model purpose, modelling choices — model structure, variable selection, resolution, and performance metrics — risk becoming disconnected from the research question they are meant to address. This disconnection underlies many of the QRPs identified in this chapter. When the link between the research question, analytic question, modelling choices, and evaluation criteria is not made explicit, it becomes easier to inadvertently select metrics that present the model favourably rather than those that genuinely assess its fitness for purpose (e.g., Fourcade et al., 2018). Seemingly minor choices in at any decision point in the modelling process can lead to different models and outcomes, and can propagate into consequential errors in management recommendations, precisely because those choices are not anchored to explicit objectives (Commander et al., 2022; Jakeman et al., 2024). The influence of analytic uncertainty underscores the need for ongoing reflexivity along the whole modelling pathway — a norm that decision-support modelling has long institutionalised but that remains underappreciated in other modelling contexts.

A second transferable norm concerns the treatment and communication of uncertainty. Decision-support modelling requires that uncertainty be characterised and communicated not as a nuisance to be minimised but as information that is integral to the decision (Dietze, 2017; Rounsevell et al., 2021). Practices such as sensitivity analysis, robustness analysis, and

scenario analysis — standard tools in decision-support — make explicit which conclusions are sensitive to modelling assumptions and which are robust to them, and thereby could help identify where analytic uncertainty and potential QRPs are most consequential. Routine adoption of these practices in basic and applied ecological modelling alike would reduce the risk of overhyping and selective reporting by making the boundaries of model credibility transparent. As Planque et al. (2022) note, assumptions that are reasonable for one model application may be inadequate for another, and modellers who do not systematically interrogate those assumptions risk drawing conclusions that do not generalise beyond the conditions under which the model was built. In both basic and applied contexts, the practice of reasoning explicitly about *what the model is for* — and ensuring that every subsequent modelling choice can be justified in relation to that purpose — is a norm from which all ecological modelling stands to benefit.

Preregistration & Registered Reports

Preregistration, and registered reports, have been hailed as a solution for preventing QRPs, and recent metaresearch empirically supports their efficacy (Burgman et al., 2023; Nakagawa et al., 2025; Purgar et al., 2024). However, there has been substantial resistance in model-related fields (Dwork et al., 2015; MacEachern & Van Zandt, 2019) because preregistration is premised on a NHST-focused definition of QRPs, that is, data-dependent analytic decisions. I argue here that there is a distinction to be made between systematic and opportunistic data-dependent analytic decisions in ecological modelling, where only the latter are questionable. For preregistration to be applied to ecological modelling, its internal logic must reflect alternative conceptualisations of QRPs — ones that accommodate legitimate data-contingent decisions, iteration, and model revision, without requiring premature commitment to a single approach (Benning et al., 2019; Evans et al., 2023; Hämäläinen & Lahtinen, 2016). In Chapter 5, I develop, apply and evaluate *Adaptive Preregistration* as a potential solution. If a complete preregistration is impractical, then at the very least, specifying a minimum set of evaluation analyses, metrics, and their performance criteria a priori is essential for avoiding QRPs.

4.6.6 Future Research

This chapter has attempted to accommodate a diversity of modelling types within ecology in its account of QRPs. However, future work should characterise how QRPs manifest across

different modelling contexts, investigating if, when, and where particular QRPs are relevant across different epistemic goals, modelling purposes, methods or research domains. To answer these questions, it would be helpful to better understand the specific incentive structures operating across different modelling contexts, beyond our focus on publication success as the primary motivating driver. Articulating the 'selection pressures' operating on models, i.e. what attributes of models are being selected for, can aid in answering questions about: which QRPs are more likely in certain modelling contexts, and whether there are specific forms they assume across contexts; and, whether some QRPs pose a greater threat to reliability than others. Decision-support modelling is a notable example of a context warranting such investigation, where stakeholder and funder pressures at the research–practice interface may function analogously to publication bias in driving QRPs. Preventative measures can then be tailored appropriately to specific contexts.

This list of QRPs is not exhaustive, future research could also characterise additional QRPs not described here, perhaps turning to other fields utilising model-based research to understand where questionable practices are more widely appreciated, such as machine learning (Garbin & Marques, 2022; Hildebrandt, 2018; Hosseini et al., 2020; McDermott et al., 2021; Meding & Hagendorff, 2024; Rosenblatt et al., 2024; Stock et al., 2023). Further, understanding the prevalence of QRPs in ecological modelling would shed light on the extent of the problem in the published literature and help prioritise potential reforms. Self-report surveys (e.g. Fraser et al., 2018) using my modelling-specific QRP classification would be a useful starting point. Empirical approaches to detecting the extent of QRPs might include approaches similar to p-curve analysis but investigating relevant model performance metrics (N. White et al., 2023).

4.6.7 Conclusion

In this paper, I aim to raise awareness among ecological modellers, and modellers among other scientific disciplines, about potential types of QRPs and their mechanisms for emergence in the modelling process. This is the first attempt to articulate how questionable research practices occur outside hypothesis testing research. The application is specific to ecological modelling, but the definition of QRPs presented here provides insights for modelling in other fields and other forms of non-hypothesis testing research. The conceptual framework and map of QRPs in this paper helps modellers understand the risks of QRPs in their research, so they are empowered to implement procedures that can mitigate their

occurrence in their own research practice. Finally, meta-researchers and advocates of open science can use the conceptual framework to underpin the design of modelling-appropriate methodological reforms that improve the credibility and robustness of model-based research in ecology and other fields.

5 "But I Can't Preregister My Research:" Improving the Reproducibility and Transparency of Ecology and Conservation with Adaptive Preregistration for Model-based Research

Preface

This chapter includes the peer-reviewed article published in *Methods in Ecology and Evolution*:

> **Gould, E.**, Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H. F., Good, M. K., Duncan, D. H., Hauser, C. E., Wintle, B. C., & Rumpff, L. (2026). 'But I can't preregister my research': Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research. Methods in Ecology and Evolution, 2041-210x.70311. https://doi.org/10.1111/2041-210x.70311.

The peer-review history for the published article is located at www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70311.

The content included in this chapter is identical to that reported in the peer-reviewed publication, however, the formatting has been altered to match the style of the thesis. Some tables have been reformatted for readability. The published article is attached in Appendix E reproduced with permission under license CC BY-NC 4.0.

Several non-traditional research outputs accompany this chapter, as follows:

- An online user-guide to the Adaptive Preregistration methodology: github.com/egouldo/EcoConsPreReg (Gould, Jones, Yen, Fraser, Wootton, Vivian, et al., 2026), reproduced in Section D.3.
- A preregistration template in multiple formats, available for installation as a quarto markdown template extension from github.com/egouldo/EcoConsModPreReg (Gould, 2026), reproduced in Section D.4.

All online supplementary material accompanying the published article is presented in Appendix D, except for Appendix S5, which is a DOCX version of the preregistration template in Section D.4.

Abstract

1. Preregistration is an open science practice which aims to improve research transparency and mitigate questionable research practices, like cherry-picking results. It helps protect against cognitive biases, like hindsight bias, that can influence how study outcomes are interpreted. There has been little uptake of preregistration in ecology and conservation, arguably because existing preregistration templates focus on null-hypothesis significance testing, whereas ecology and conservation often rely on different types of statistical modelling.
2. We argue that preregistration in model-based research in ecology and conservation is both possible and beneficial, using templates adapted for domain-specific methodologies. We applied a user-centred design approach to translate the concept of preregistration into model-based research practice for ecology and conservation.
3. To better align the internal logic of preregistration with the iterative and non-linear process of ecological modelling, we propose, test and evaluate a methodology for ‘Adaptive Preregistration’, using a case study of modelling managed water releases (‘environmental flows modelling’) in regulated rivers for maintaining riparian vegetation condition in Victoria, Australia.
4. This research provides a template and methodology for implementing Adaptive Preregistration of ecological models. Although we focus on ecology and conservation in this paper, the concept of Adaptive Preregistration, and the templates developed here, could be applied to model-based research in other scientific disciplines. Modellers in ecology and conservation need no longer cry ‘but I can’t preregister my research’.

Keywords: ecological modelling, environmental flows modelling, good modelling practice, metascience, open science, preregistration, reproducibility, transparency.

5.1 Introduction

Over the past decade a 'replication crisis' in some scientific fields has exposed problems with the reliability of research findings (Camerer et al., 2018; Neoh et al., 2023). One such problem is the prevalence of questionable research practices (QRPs), including *p*-hacking, hypothesising after results are known (HARKing), and cherry-picking (Wicherts et al., 2016). QRPs increase the probability of false positive results, artificially inflate effect size estimates, and ultimately reduce the replicability of published findings, while encouraging overconfidence in the precision of results and exacerbating the risk of accepting and propagating false facts (Hoffmann et al., 2021). Researchers' self-reports show that QRPs are surprisingly common across disciplines (Agnoli et al., 2017; Liu et al., 2020; Makel et al., 2021), including in ecology (Fraser et al., 2018; Kimmel et al., 2023). Although self-report surveys of QRPs are heavily focused on statistical significance testing research employing *p*-values, some practices, like selective reporting/cherry-picking, occur and operate analogously in other inferential frameworks.

QRPs are precipitated when researchers opportunistically or systematically misuse 'researcher degrees of freedom' without disclosure (Bakker et al., 2020; Wicherts et al., 2016). Researcher degrees of freedom refer to the many decisions and alternative analytic choices researchers make throughout the research process (Wicherts et al., 2016). Analytic decisions are typically subjective but methodologically defensible; however, choices and omissions taken at each 'fork' may collectively and significantly influence the final 'analysis strategy' (*sensu* Hoffmann et al., 2021), or 'modelling path' (*sensu* Hämäläinen & Lahtinen, 2016). The set of all (plausible or reasonable) possible decisions that a researcher could make over the entire analysis from beginning to end has been termed collectively as the 'garden of forking paths' (Gelman & Loken, 2013). 'Many-analyst studies', where multiple researchers independently investigate the same research question by analysing the same dataset, have now demonstrated the consequences of the 'garden of forking paths,' whereby alternative analytical decisions can lead researchers to make vastly different interpretations of results, with conflicting conclusions (Silberzahn et al. (2018), and Chapter 2). QRPs emerge when researchers opportunistically navigate this garden through systematic selection bias, tending

to follow and report on only the analytical paths that yield the cleanest, strongest or most favourable outcomes, while leaving alternative paths unexplored or undisclosed.

Recent frameworks for understanding QRPs in ecological modelling show that these practices involve artificially inflating the apparent credibility, accuracy or performance of models through this opportunistic path selection (Chapter 4). Modellers may engage in 'model fishing' by trialling multiple model structures (e.g. linear, polynomial, mechanistic) and reporting only those with the best fit statistics; 'S-hacking' by testing various performance metrics (e.g. R^2, AIC, accuracy thresholds) and selectively reporting only favourable results; or 'sample curation' by removing inconvenient sites or time periods that do not support expected patterns (Chapter 4). When this selective path-taking is repeated across studies, it contributes to publication bias, where negative or inconclusive results remain unpublished in the file drawer, culminating in a distorted literature that masks the true uncertainty and variability in ecological systems while making published models appear more credible and reliable than they actually are.

Good intentions and simply being aware of the potential for researcher degrees of freedom to lead to QRPs cannot sufficiently mitigate their risk because of the unconscious nature of their origin (Zvereva & Kozlov, 2021). Preregistration is an open science practice that aims to distinguish between planned and ad hoc analyses (Parker et al., 2019), thereby restricting opportunities for trialling analytical alternatives ('researcher degrees of freedom') and making any deviations from the preregistered plan transparent (Wicherts et al., 2016). Preregistration requires researchers to register their methods and analysis plan in a secure and publicly accessible platform prior to collecting and/or analysing data, which cannot be altered after submission (Parker et al., 2019). Preliminary empirical evidence shows that studies with preregistered analysis plans are more transparently reported and are less likely to contain positive or significant results (Brodeur et al., 2024). This finding in part reflects the researcher commitment to honest, complete and transparent reporting, especially when the preregistration template includes detailed instructions and requirements (Bakker et al., 2020), and a correction for the selective reporting practices that contribute to publication bias. Preregistering studies can also combat the impact of publication bias by making studies findable regardless of whether they are published, in a similar way to a medical trial registry. 'Registered Reports', a publication format that integrates preregistration into the peer-review process (Koivisto & Mäntylä, 2024), contain a direct mechanism for addressing

publication bias, whereby the methods and analysis plan are peer-reviewed in advance and the decision to publish is made *before* the results are known.

One potential objection to preregistration for modelling is that model development is an intrinsically exploratory process (MacEachern & Van Zandt, 2019), while preregistration is for confirmatory research (i.e., hypothesis or theory testing, Prosperi et al., 2019). We acknowledge that preregistration serves a crucial role in documenting research intentions and preventing HARKing and other QRPs, and that the exploratory/confirmatory distinction provides valuable guidance where it applies clearly. For model-based research that can be fully specified in advance—where analytical strategies are well-established and data-dependent decisions are minimal—'static' preregistration remains the preferred approach. These conceptions of modelling and preregistration are not entirely wrong, but they are oversimplified for ecological modelling. In practice, there is not such a clear distinction between exploratory and confirmatory activities (Alspaugh et al., 2019; Devezer et al., 2021; Flournoy et al., 2020). Modelling includes some tasks and activities that are exploratory in nature but can be distinguished from wholly exploratory analysis in that modelling is often goal-driven and directed (Alspaugh et al., 2019). Researchers move back and forth between different types of reasoning and inference throughout the modelling process, with deductions leading to abductions as the analysis proceeds (Alspaugh et al., 2019; Prosperi et al., 2019) or even occurring together in a hybridised manner (Prosperi et al., 2019).

In this iterative and non-linear process of modelling, initial analyses often inform future analyses of the same data, whereby the modeller's understanding of the problem and the model itself are incrementally adjusted in light of intermediate results from previous decision points (Dwork et al., 2015; Hämäläinen & Lahtinen, 2016). Recent evidence from many-analyst studies supports this view: Schweinsberg et al. (2021) found that even within individual analyses, researchers engage in both exploratory and confirmatory activities iteratively, contrasting against definitions of 'confirmatory' analysis used within the preregistration literature, where confirmatory analyses are typically 'planned out and 'frozen' online prior to having the dataset'. Modellers are typically uncertain about which model is most appropriate for the data (Popovic et al., 2024) and are reluctant to register a modelling strategy that may turn out to be incompatible with the data (Roettger, 2019). For example, modellers may perform exploratory analyses to validate distributional assumptions (Campbell, 2021), which may result in subsequent changes to the model specification. We argue that systematic data-dependent decisions like these should not be interpreted as

violating a key tenet of preregistration, that is, failing to define the entire analysis protocol ahead of time (Dwork et al., 2015; MacEachern & Van Zandt, 2019). The key distinction is not whether decisions are data-dependent, but whether they are systematic (following preregistered decision rules) versus opportunistic (result-seeking without disclosure, see Chapter 4).

We propose a methodology for implementing an expanded view of preregistration called 'Adaptive Preregistration' (*sensu* Srivastava, 2018), which captures decision points and reasoning about modelling choices, and embraces the data-dependency inherent within the modelling development process for research contexts that don't fit neatly into binary exploratory/confirmatory classifications. This goes beyond the work of others who offer discussion and guidance about when and how to deviate from a preregistration (Lakens, 2024; Willroth & Atherton, 2024), instead providing an approach to preregistration embodying the principles of 'registered flexibility' (Roettger, 2019; Srivastava, 2018) and iterative or 'interim preregistrations' (Hofman et al., 2023; Ioannidis, 2022).

In this study, we aim to translate preregistration for application to model-based research, specifically aimed at ecology, conservation and related disciplines. Here, we first identify a generalised 'modelling workflow' that captures critical analysis decisions in the model development cycle and use it to build an Adaptive Preregistration template that can work for ecological modelling. We then test and evaluate our template using a case study of modelling managed water releases in regulated rivers ('environmental flows') to improve riparian vegetation condition in Victoria, Australia. This case study demonstrates a novel approach to preregistration using a template designed for model development and evaluation.

5.2 Materials & Methods

We modified preregistration for use in model-based research contexts following a user-centred design framework (Pavelin et al., 2012), where users were engaged at all stages to produce a preregistration template and protocol that is fit for purpose (Figure 5.1). All study materials are archived at osf.io/sq5j4 (see Appendices D.1 and D.2).

5.2.1 Problem Formulation: Defining vision and objectives for a preregistration template for ecological modelling

We first specified the research context, purpose, target users and use-cases (see Figure 5.1, step 1; Pavelin et al. (2012); Pu et al. (2019)). Our focus is model-based research, which we define broadly as any research that uses quantitative modelling to answer its research question (Brudvig, 2017). This encompasses diverse modelling goals (exploration, inference, and both explanatory and anticipatory prediction, Bodner et al., 2020; Mouquet et al., 2015; Tredennick et al., 2021), varied research contributions (developing new methods, applying or extending existing approaches, and model transfers; Yates et al. (2018)) and different model types (phenomenological vs. mechanistic, Connolly et al., 2017) across basic and applied research contexts and disciplines (Brudvig, 2017; Lewis et al., 2023). We intentionally maintain this broad definition because ecological modelling is highly diverse in its methodological approaches, analytical purposes and research applications.

The purposes of the preregistration template and methodology are to:

1. Delimit 'researcher degrees of freedom' (Pu et al., 2019) to mitigate the risk of QRPs when conducting model-based research in applied ecology and conservation.
2. Increase 'research transparency' of ecological modelling*, including both:
    i. *production transparency* – including research artefacts like open-access data and materials, or data collection procedures (Lupia & Elman, 2014),
    ii. *analytic transparency* – a complete account of how analytic conclusions are drawn from the data (Lupia & Elman, 2014) where model choices, steps in the modelling process, assumptions and expectations about model outputs are clearly articulated (Bodner et al., 2020).

We specified two use-cases for our Adaptive Preregistration template:

1. Study authors or modellers seeking to preregister their modelling study.
2. Editors, reviewers, and/or readers of the completed preregistration, who are tasked with 'preregistration checking' or verifying that analyses and findings reported in the final manuscript are consistent with the preregistered analysis plan (Pu et al., 2019).

Existing preregistration formats are designed primarily for study authors; they are static text-based documents that do not easily facilitate preregistration checking. We designed our template with editors, reviewers and readers in mind.

5.2.2 Designing the General Modelling Workflow

Building on our broad definition of model-based research (Section 5.2.1), we conducted a literature review (Appendix D.1) to identify a modelling workflow suitable for application to a range of different modelling goals, research contributions, model types and problem contexts. The final output was a workflow that reflected a balance between idealised and actual modelling in practice, aligned with the prescriptive guidelines of 'good modelling practice' (Augusiak et al., 2014).

5.2.3 Designing the Preregistration Template

We organised and facilitated a workshop with 10 ecologists and ecological modellers (Appendix D.2). A key task was to identify common modelling workflows and critical decision-steps in model development (Figure 5.1, step 2). 'Critical' decision-steps are points in the modelling process where researchers, or modellers, make analytical decisions that could change the output of the analysis. In this activity, workshop participants first individually reflected on their own scientific process for a recent or memorable research project, listing the steps undertaken throughout the process. Next, participants collaboratively mapped the decision steps in their own personal modelling workflows onto a modelling workflow template prepared earlier (osf.io/fgd23), informed by our general workflow (Figure 5.1). We then collaboratively reviewed and refined the modelling workflow into a standard set of modelling phases and commonly implemented decision steps that formed the basis of the preregistration template (Figure 5.1, step 3).

5.2.4 Developing an Adaptive Preregistration Methodology

We developed a methodology for application of preregistration to model-based research, or Adaptive Preregistration (Figure 5.1, step 3), comprising two key components: registered flexibility and interim preregistrations. We summarise the methodology and direct readers to our user-guide for more detailed guidance on implementing Adaptive Preregistration in practice (Appendix D.3).

Registered Flexibility

Registered flexibility includes preregistration of 'plans to deploy flexible strategies' (Srivastava, 2018), whereby the preregistration author specifies flexible heuristics containing alternate analysis or modelling strategies whose execution depends on the outcome of previous decision-points or analyses. For example, when faced with methodological or analytic uncertainty that cannot be resolved without observing parts of the data, conducting preliminary analyses, model checking or other data-dependent decisions, a modeller can preregister a decision-tree that consists of predefined rules about when a particular modelling strategy or decision should be implemented (Baldwin et al., 2022). There are three requirements for adequately registering flexible analyses:

1. Stating what quantity needs to be known to move forward with the modelling and analysis, and why.
2. Describing the analysis that will be used to generate this quantity and which parts of the data will be used.
3. Explaining how the results will be interpreted, listing each alternative decision under consideration and the analysis result that will trigger each decision, respectively.

Interim Preregistrations

The modeller follows an iterative process of preregistration in parallel with modelling, consisting of interim preregistrations that mark key phases of modelling and analysis as different parts of the data are observed (Srivastava, 2018). As the modeller proceeds through the model development process, they shift from ideation and preregistration to execution of the preregistered analysis plan, and back again, generating interim preregistrations at different points within the model development process, depending on observed outcomes of the flexible heuristics described in the interim preregistration (Figure 5.2). The realised modelling path executed from the garden of forking modelling paths will depend on the observed outcomes of the registered flexible analyses at each interim preregistration.

Transparent Documentation of Adaptive Preregistration

Ideally, to transparently document the Adaptive Preregistration process and facilitate preregistration checking of flexible modelling and analysis strategies (one intended use-case for our methodology and template, Section 5.2.1), the results of any interim and final analyses

must be explicitly linked back to the preregistered analysis strategy. Researchers can take simple steps to achieve this, for example, by retaining separate versions of initial, interim and final preregistration documents; retaining the implementation and results of registered flexible analyses alongside the preregistration documents while recording their filenames in the relevant preregistration document; and utilising existing platforms with basic version control features, such as the Open Science Framework (OSF, cos.io/products/osf).

Optional: Preregistration with git & GitHub

The use of version-control for transparently documenting code-based analyses is considered best-practice, however, is not yet widespread in ecology and related disciplines (Braga et al., 2023). We propose using git and GitHub as the vehicle for transparent documentation of Adaptive Preregistration of a modelling study for those researchers already comfortable with these tools (see user guide for details, Appendix D.3). The procedure leverages GitHub's tag and release feature (help.github.com/en/github/administeringarepository/releasingprojectsongithub) in conjunction with Semantic Versioning (semver.org) to track, document and collaborate on changes to the preregistration document and the analysis. Modelling code and results are stored and versioned alongside the preregistration document in the project repository, while GitHub 'issues' (see Braga et al., 2023 for explanation and definition) are used to track discrete analysis tasks described in the preregistration. GitHub issues are particularly useful for registered flexible analyses, facilitating documentation of any interpretation of the results of registered flexible analyses and their influence on the rationale for the subsequent preregistration; the outcome of the registered flexible analysis is recorded within that issue, hyperlinked to the registered heuristic in the preregistration document, alongside interpretation of those results and documented decisions about which decision alternative is to be selected based on those results.

5.2.5 Evaluating and Refining Adaptive Preregistration with a Case Study

We preregistered a real-world model-based research problem as a case study for evaluating the preregistration template and proposed Adaptive Preregistration methodology (Figure 5.1, step 4). The case study analysed the effectiveness of environmental flows on maintaining riparian vegetation condition in Victoria, Australia (Jones et al., 2025; Jones, Vivian, et al., 2026), forming a component of the Victorian Environmental Flows Monitoring and Assessment Program (VEFMAP), a large-scale, long-term monitoring programme

delivered by the Arthur Rylah Institute for Environmental Research. Case study data were collected under a research permit for the Victorian Flora and Fauna Guarantee Act, permit number 10010953. The VEFMAP research team consisted of a project lead (CJ), who was responsible for establishing the conceptual framework and research questions that inform the analysis, and a modeller (HW) who undertook much of the coding and exploratory analysis under the direction of a lead modeller (JDY). Modelling experience among the team ranged from moderate to advanced, and the preregistration was collaboratively completed by the case study research team and the study lead (EG). In addition to following the protocol for Adaptive Preregistration using git and GitHub, we also utilised the version-control and collaborative project management features of git and GitHub (Braga et al., 2023) to live–develop the preregistration template and capture the case study researchers’ feedback about the template and Adaptive Preregistration methodology. For example, if it emerged in the process of completing the case study preregistration that we had missed an important step in the modelling process or found that the order and structure of the template should change, the suggested change and its justification was recorded in GitHub discussions. We also conducted follow-up semi-structured interviews with the case study researchers to capture detailed reflections.

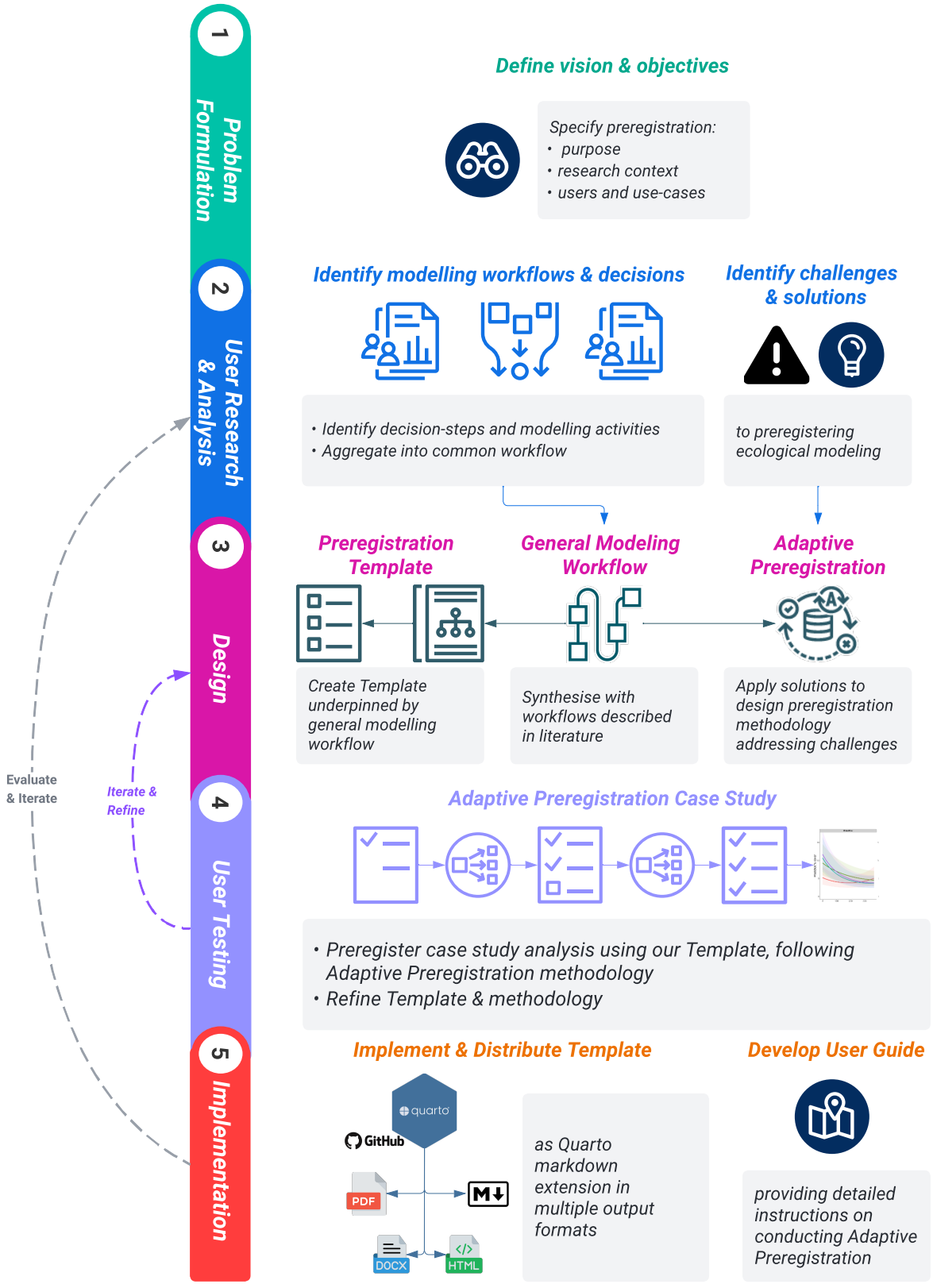


Figure 5.1: *Process for translating preregistration to ecological modelling research. We followed the user-centred design steps proposed by Pavelin et al. (2012), except for the ‘Evaluate and Iterate’ phase of the process, which we hope will be taken up by the community when trialling the practical tools developed in this study. Following a collaborative workshop (Step 2), we specified a general modelling workflow that informed the structure and content of a preregistration template (Step 3) and proposed a methodology we call ‘Adaptive Preregistration’ (Step 3) addressing challenges to preregistering model-based research identified in Step 2. We refined the template and methodology using a real-world modelling case-study (Step 4). The template is provided in multiple formats (Step 5), including as a Quarto Markdown extension (Gould, 2026), and instructions for conducting Adaptive Preregistration are detailed in a user guide (see Appendix D.3).*

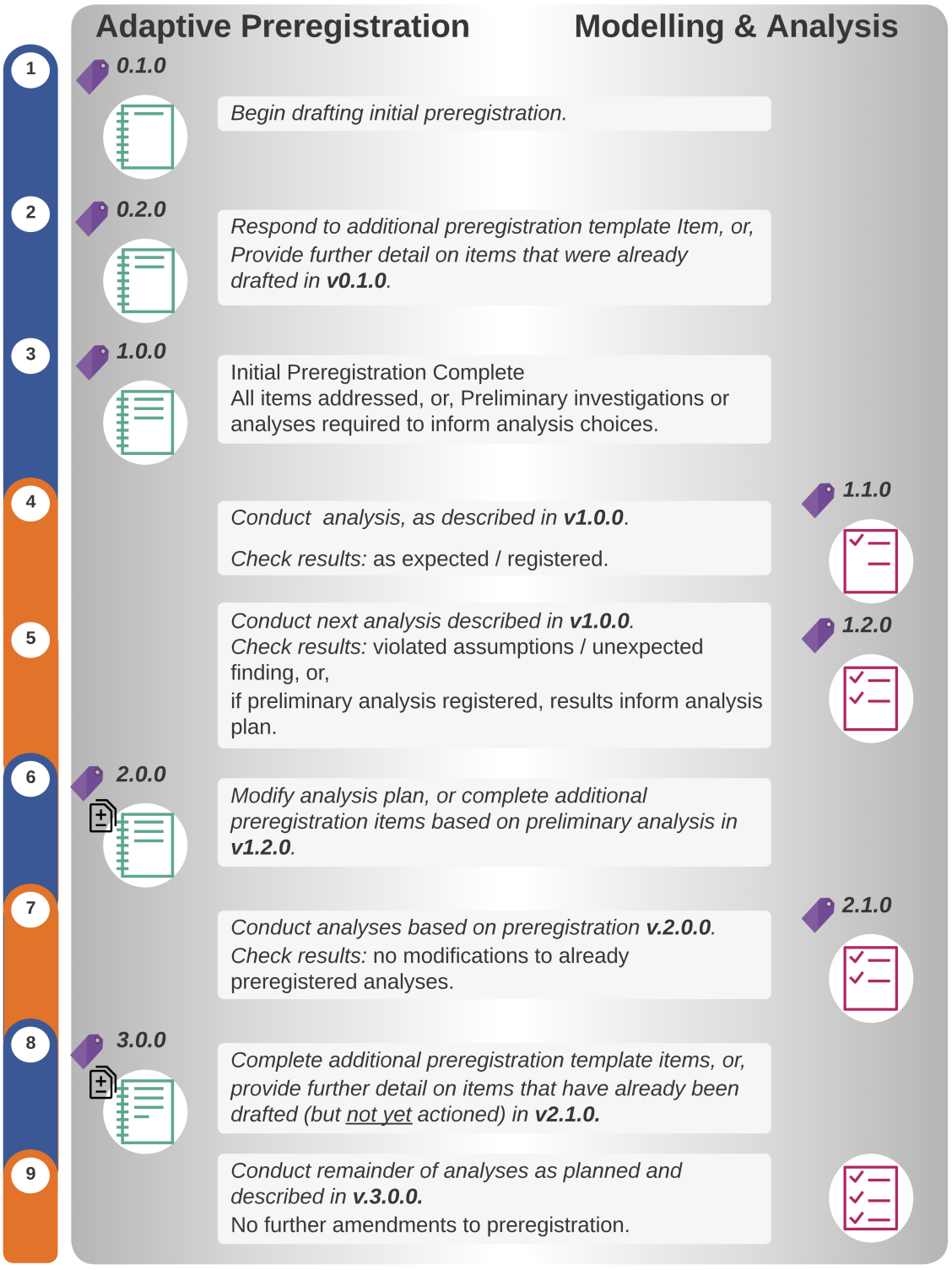


Figure 5.2: *Implementing Adaptive Preregistration. The preregistration (blue steps) and modelling and analysis processes (orange steps) operate in parallel. Purple tags represent git tags, corresponding to incremental versions of the preregistration. Any changes to the analysis plan receive a GitHub release and major version number increment. Step 3 corresponds to the initial preregistration document, Step 6 corresponds to an interim preregistration and Step 8 corresponds to the final preregistration. The execution of the Final Preregistration occurs in Step 9. Note that the use of git and GitHub are optional (see Section 5.2.4, p. 144).*

5.3 Results

5.3.1 The Preregistration Template

The final preregistration template, refined through application of the draft template developed in the user-research workshops to the case study, is hosted at github.com/egouldo/EcoConsPreReg and provided as a Quarto Markdown extension (Gould, 2026) that can be installed and rendered in HTML, PDF and DOCX formats (Section D.4). Intermediate versions of the template developed through application of the initial template to the case study are available in the case study preregistration document's GitHub commit history (github.com/egouldo/VEFMAP_VEG_Stage6/commits/master/analysis/preregistration_template.Rmd, osf.io/sq5j4/files/aup5f) or in case-study GitHub repository releases (github.com/egouldo/VEFMAP_VEG_Stage6/releases).

5.3.2 Findings from the Case Study

The application of our preregistration template and Adaptive Preregistration methodology to the environmental flows modelling case study occurred across three distinct stages (Table 5.1, Figure 5.3), with two interim preregistrations (V1 and V2) preceding the finalised preregistration (V3, Appendix D.5; Jones, Yen, et al. (2026)), each corresponding to distinct phases in both the 'Preregistration' and 'Modelling and Analysis' processes. While preparing Version 1, the modellers decided that they needed to conduct analyses to inform specification of appropriate candidate models. They preregistered a 'Preliminary Analysis' (V2) which used a subset of the full dataset (which was not discarded) to preserve some degree of data-dependent decision-making while permitting a degree of exploratory analysis. The Preliminary Analysis involved exploratory analyses to inform model specification and analysis for the 'Main Analysis' on the full dataset (V3, Appendix D.5; Jones, Yen, et al. (2026)). Several classes of data-dependent decisions (Liu et al., 2020) emerged during our case study, which were managed using a combination of registered flexibility and interim preregistrations:

- *Dataflow dependencies*: where the output of one decision is the input to another. For example, the models parameterised on the full dataset in the case study's Main Analysis were subject to a suite of model checks as outlined in the decision-trees in

Table 5.1: *Key stages of the modelling and Adaptive Preregistration process implemented in the environmental flows modelling case study. Version numbers in parentheses are GitHub tags, marking snapshots of the repository for each Preregistration Version (github.com/egouldo/VEFMAP_VEG_Stage6/releases, osf.io/sq5j4/files/jdbr7).*

Preregistration version	Preregistration process	Modelling and analysis process
1: (v0.1.0 - v0.8.1)	Establish problem context, study background and aims, describe data collection and data cleaning process, articulate modelling objectives and specify candidate models	Data cleaning and preparation undertaken concurrently by lead and supporting researchers due to limited resourcing and short timelines for project deliverables. Candidate models addressing different components of the research question were preregistered. Data structural properties and potential consequences for model fitting were identified
2: (v0.9.0 - 0.12.5)	Describe Preliminary Analysis on data subset and specify registered flexibility in the form of decision heuristics	Critical uncertainties in modelling decisions were articulated (e.g. how do we classify best flow regime based on inundation data?). Additionally, data partitioning, exploratory data analysis, refined candidate models and model checking procedure were preregistered
3: (v0.13.0 - 0.20.1)	Describe Main Analysis, including specification of decision tree for model selection process based on results of Preliminary Analysis	Conducted modelling and analysis as outlined in preregistration. Reported deviations and rationale as necessary

Preregistration Version 3 (Figure 5.4), with the results of model checking informing a priori specified alternative model structures and model functional forms.

- *Information dependency*: where one decision informs another. For example, outcomes from the exploratory Preliminary Analysis (Interim Preregistration V2, Table 5.1) informed the model structures specified in Preregistration Version 3 (Table 5.1), as well as the decision heuristics for triggering their fitting and acceptance.
- *Procedural dependency* : where downstream decisions would not exist if some alternative upstream decision was made instead. For example, if the Preliminary Analysis had not revealed over-dispersion or zero-inflation in the data, then the case study modellers would not have fitted zero-inflated Poisson models, nor would they have trialled accounting for over-dispersion or refitting using negative binomial models (Figure 5.4).

Interim Preregistration (Versions 1 and 2)

Two initial candidate 'maximal' models were preregistered in Version 1 (Figure 5.3), one for each study response variable; however, modellers were unsure if these models would

be supported by the dataset due to low sample sizes relative to the complexity and spatiotemporal patchiness of the data. Preregistration Version 2 described an exploratory Preliminary Analysis that aimed to resolve critical uncertainties about how to operationalise the models, given data constraints and how to specify a candidate set of models that could feasibly be fitted to the full dataset (Table 5.1). This involved partitioning the dataset into a ‘Preliminary Analysis’ subset for making data-informed modelling decisions, which were then preregistered in a subsequent preregistration of the Main Analysis (V3, Table 5.1; Figure 5.3). The Preliminary Analysis dataset consisted of observations from a single site with good temporal data coverage, while the Main Analysis dataset comprised the full dataset, including the Preliminary Analysis dataset. Preliminary Analysis informed how to operationalise the models given the underlying structure of the dataset. For example, simple tests of data distributions within hierarchies were required to check for data spread and the presence of zero-inflation, while checks for collinearity among candidate predictor variables were needed to prevent confounding and overfitting, and to ensure relevant interactions were captured in the models. The same procedures for evaluating model fit and model checking were preregistered for both the Preliminary and Main Analysis (Figure 5.4 b), however, the purpose of model checking in the Preliminary Analysis was to identify appropriate model structures, whereas the purpose for the Main Analysis was to ensure that the fitted models were supported by the data.

Final Preregistration (Version 3): Main analysis

In Version 3 of the Preregistration, we derived and preregistered:

1. A preferred candidate model or ‘full model’ for each response variable (vegetation richness and vegetation cover), to be fitted to the full dataset (Initial Preregistration, Figure 5.3; Appendix D.5, p. 148; Jones, Yen, et al. (2026)).
2. A candidate set of simplified models that separately captured different flow components for each response variable, with two versions of the flow events model for vegetation cover (Flow Events Models A & B, Figure 5.3; Appendix D.5, pp. 155 - 156; Jones, Yen, et al. (2026)).
3. Decision-trees for triggering fitting and acceptance of the candidate simplified models should the preferred full model specifications fail to converge and/or provide appropriate parameter estimates (Figure 5.3, Figure 5.4

a. A second flexible strategy, for determining model functional form and to account for potential over-dispersion (Figure 5.3; Figure 5.4 b). Preregistering this flexibility was motivated by the modellers’ concerns that the actual data distributions in the full dataset did not match expectations informed by the Preliminary Analysis.

This final decision-point for choosing the model family is subjective and fuzzy, and explicit weightings between the three model performance criteria could not be expressed or preregistered. Instead, modellers explained that the final decision will be guided by the model’s overall ability to capture key associations reliably.

5.3.3 Adaptive Preregistration in Practice: Benefits and challenges

At the end of user-testing, case study researchers concluded that preregistration is a valuable tool to facilitate good modelling practice, but also acknowledged several difficulties encountered during Adaptive Preregistration (Table 5.2).

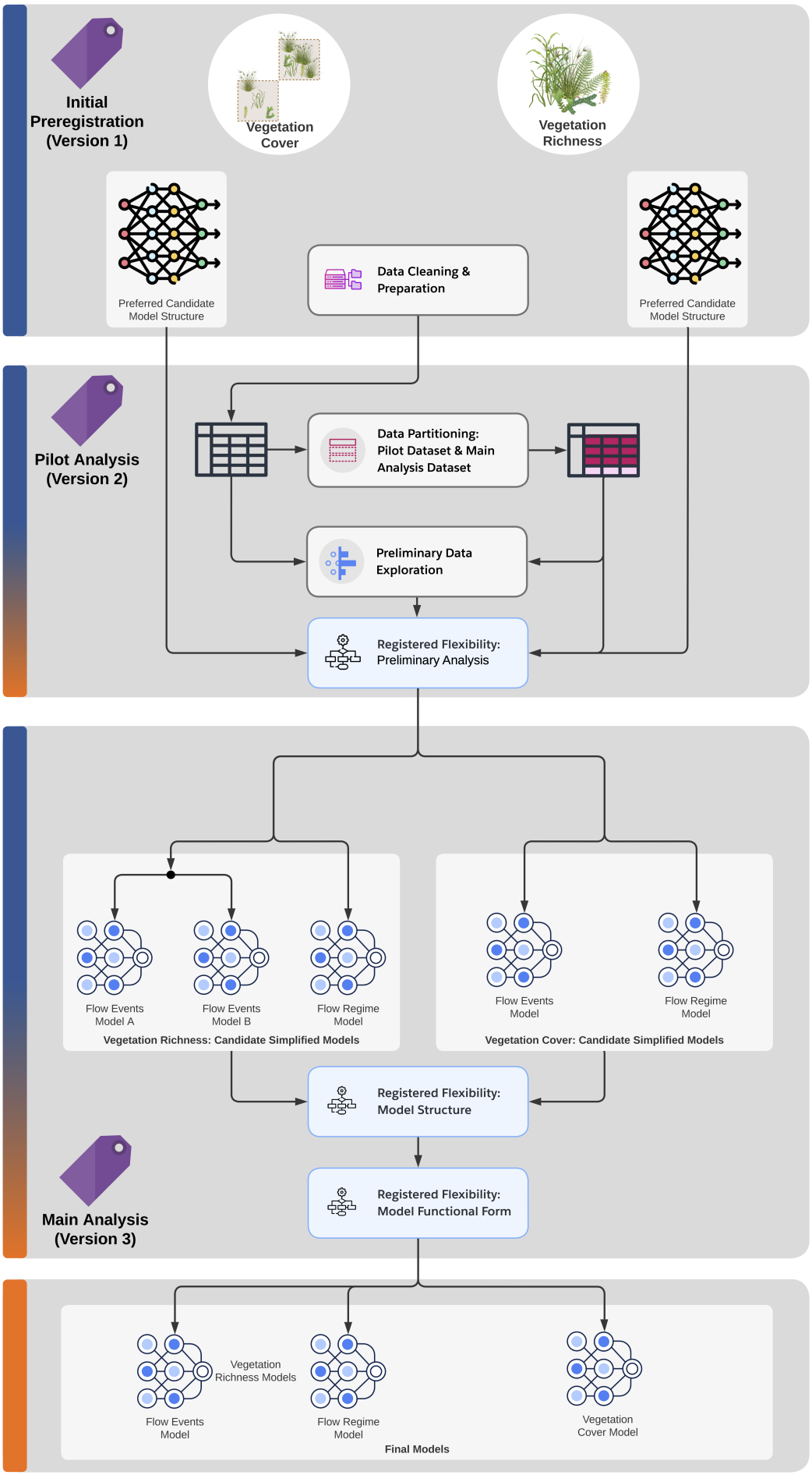


Figure 5.3: *Evolution of model development across preregistration iterations in the case study. Each preregistration version corresponds to different stages of model development, summarised in Table 5.1. Maximal models for each response variable were specified in the initial preregistration (V1). Exploratory and preliminary analyses and data partitioning were next preregistered (V2), aiming to resolve uncertainties in model specification given data constraints. The Preliminary Analysis identified three candidate simplified models for vegetation richness and two for vegetation cover. These model structures were preregistered in the final preregistration for the Main Analysis (V3), in conjunction with registered flexibility for determining model structure and functional form. The two registered flexibility steps in the Main Analysis (shown in blue) were informed by data exploration and registered analyses conducted during Preliminary Analysis (V2) and are further described in Figure 5.4. The Main Analysis was preregistered to be conducted on the full dataset. While conducting the preregistered Main Analysis, three final models were identified for model fitting and analysis, one model for vegetation cover and two vegetation richness models, with one incorporating ‘flow events’ (spring and summer freshes) and the other incorporating ‘flow regime’ (days above baseflow and days above spring fresh [per year]).*

Table 5.2: *Summary of challenges ⊖, benefits and added value ⊕, as well as tips for using Adaptive Preregistration for ecological modelling.*

Type	Name and description	Case study example	Practitioner recommendations and reflections
	Decide on template resolution Time required to make preregistration templates specific, precise, and exhaustive may be off-putting. What is an appropriate resolution?	The preregistration template first used in the case study was unwieldy and overwhelming. Resolution was reduced by restructuring, removing and merging template items, making the task more approachable	*'Be willing to modify the template if it does not meet the needs of the project. Any attempt to preregister is better than none at all'* — JDY
	Decide on template formatting The visual design, presentation and formatting affect usability	Reformatting the template with collapsible help and explanatory text boxes, as well as coloured icons to differentiate preregistration items from explanatory text, improved clarity of the initial preregistration and facilitated more efficient completion of the preregistration using the template	
⊖	**Choose a version-control platform that does not require too much investment in upskilling** Adaptive Preregistration requires interim versions of the preregistration be deposited in a single location that is ideally version controlled. Some platforms have a steeper learning curve than others, impacting on the efficiency of preregistration	Using GitHub to track iterations in the preregistration proved difficult to combine with a GitFlow approach for simultaneously collaborating on the analysis code and preregistration across branches. The lead modeller (JDY) regularly uses git for version control but was less familiar with the GitHub web GUI. Under broad direction of the lead modeller, a second modeller (HW) undertook much of the coding and exploratory analysis, but had limited familiarity with version-control systems, and was unfamiliar with collaborative code-base development tools, such as GitHub. Substantial project time was allocated to developing a collaborative GitHub workflow and familiarising both modellers with this approach	
	Ensure that modelling decisions and registered flexibility are accurately documented and easily discernible when preregistration checking For complex modelling studies, especially where there are multiple models being preregistered and/or where there are multiple models under consideration within a model selection approach, it is difficult to keep track and accurately document analysis decisions, including flexible decision-heuristics	When checking the preregistration to write the case study report, the lead author (EG) found it difficult to clearly identify registered decision-triggers for choosing alternative model functional forms over preferred forms, due to relevant information being dispersed over multiple subsections within the preregistration (Appendix D.5, 4.3.1 Quantitative Model Checking, 3.2.2 Choose Model Family, 3.5 Model Assumptions and Uncertainties, 3.4.2 Estimation Performance Criteria; Jones, Yen, et al. (2026)). Moreover, some decisions concerning the model selection process were informed by tacit knowledge of the modelling team and not recorded in the preregistration (Section 5.4.2)	*'I could not figure out why quadratic associations rather than linear associations were fitted at the functional group level when these were not described in the preregistered full model descriptions and pseudo-code. After significant effort, it became apparent that this decision was described across multiple subsections'* — EG. *'In hindsight, we would specify the model selection procedure in both the Preliminary Analysis and Main Analysis more clearly. The process we followed was more akin to "make the model more complex again, and make it look more like the full model"'* — JDY

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊖ 	**Adaptive Preregistration may be incompatible with model management workflows in real world applications** Operational procedures may constrain analytic decision-making too early in the research timeline and require modellers to be involved earlier and throughout the research. Operational constraints may be easier to overcome in academic rather than government or industry settings	At the case study institution, similar to many organisations working on applied problems, modellers are typically brought into the research project late in the research programme and are allocated limited time over a discrete period (e.g. 5 days immediately before report-writing and well after project planning) to conduct the modelling with limited knowledge of data provenance, dependency structure and underlying distributions. Following the Adaptive Preregistration procedure — where there were several rounds of iterative planning and preregistration, and implementation and analysis — required modellers to be involved much earlier in the research process than usual. Changes in operational procedure and culture would need to occur at the case study institution to facilitate Adaptive Preregistration becoming commonplace	
⊖	**Preregistering flexible analyses is difficult and may delay the research process, especially as the analysis increases in complexity** Formulating alternative decision-pathways and their decision-triggers is challenging under certain conditions: when the underlying dataset is complex, unfamiliar to the modellers, zero-inflated, over-dispersed and small (low sample size). On the other hand, preregistration of straightforward projects would likely take less time due to increased familiarity with 'standard' or 'go-to' analyses for simpler data types and structures	The process of working through and documenting analysis decisions prior to analysing the data delayed completion of the analysis considerably when compared to expectation. The task of mentally forecasting potential analysis scenarios and decision pathways was particularly challenging in this instance, because the dataset was complex and unfamiliar to the modellers. Compared to knowledge and models of processes affecting fish responses to environmental flows, the vegetation models are relatively less developed. The lead modeller has much more familiarity modelling count data on fish and reflected that there would be less analytic unknowns in those instances, which would have made the task of preregistering flexibility much easier	*'Adaptive Preregistration did delay the research process initially, because we had to carefully consider all the likely scenarios to the end of the process and document those, as opposed to only needing to figure out the first part and then starting. I expected this delay to be about a month; however, the delays were much longer due to the fact that we had never tried preregistration before, we were also feeding back information into the Adaptive Preregistration process, and the model design and structure was very complex and different to things we had done in the past. It was a pretty challenging analysis — far more complex than many ecological studies I have worked on'* — CJ

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕⊖	**Adaptive Preregistration is difficult at first and therefore slows the research process** but will become easier and more efficient with familiarity and experience with preregistration	The preregistration of decision trees and heuristics for triggering the selection of some models over others was the most difficult and time-consuming aspect of Adaptive Preregistration for the case study modellers. However, with practice and experience they expect Adaptive Preregistration will become easier and quicker	*'The process would definitely be easier and faster the next time, even for a project of the same complexity. Partly this is because of just knowing how it works and what to expect. The first one will always be the most difficult and slow. Going through the Adaptive Preregistration process once or twice will make it much more efficient'* — CJ
⊕	**Planning the analysis at the outset likely saves time in the long run** Using a structured workflow that carefully considers modelling decision points and options in advance reduces scope for time-consuming back-and-forth adjustments later in the process. Having a documented plan to refer back to also saves time on similar, future tasks and report writing	Typically, the modelling team follows a more trial-and-error-based approach to model development, particularly for research and decision-making contexts where conceptual models, variable definitions and model parameters are uncertain (such as for the vegetation components of VEFMAP, compared to fish models). Adaptive Preregistration forced the modellers away from their usual iterative process towards carefully considered and well-justified a priori construction of the models. The case study modellers thought that the modelling process saved time in the coding of the models and by preventing iterative adjustments in the models and code and saving potential back-and-forth communications between modelling team members when deciding on and implementing those changes	*'Undoubtedly using Adaptive Preregistration saved some time at the end of the process, but it is hard to say how much. Shifting the process from a more trial-and-error based approach to a very carefully considered construction of the model a priori, must surely have saved time in the coding and back-and-forth adjustments stages. The document was very helpful when it came to write the report. We also have a worked example that we are happy with that we can use to leverage off or remind ourselves how we approached a step in the past'* — CJ

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕	**Preregistering model structures and predictors in advance strengthens links between methods and the original study aims/concepts**	Preregistration required careful consideration of the candidate predictors and possible model structures. The case study modellers acknowledged that while the broad approach to model development would likely have followed a similar progression (i.e. expanding on a basic Poisson model to address zero-inflation and over-dispersion, while also reducing a candidate predictor set down to a maximal viable set), in the absence of Adaptive Preregistration, they may have proceeded without clear criteria for selecting one model over another (noting that these criteria were still relatively vague in the preregistered study)	*'If these models were developed without preregistration, much of this thinking would have been "on the fly," which may have resulted in better model fit (under a range of criteria) but perhaps weaker links to the study's aims and conceptual basis'* — JDY
⊕	**Adaptive Preregistration reduces researcher degrees of freedom** by shifting the nature of the modelling process away from a loosely bounded and iterative, trial-and-error style strategy towards a more constrained, goal-directed analysis. Incorporating loosely defined preliminary analyses with data subsetting facilitates data exploration being formally incorporated into the analytic decision-making process	When considering how the case study modelling would have occurred without being preregistered, both the project lead and lead modeller thought that the preregistration process led to similar outcomes to what would be expected under standard practices, but that the modelling process under Adaptive Preregistration may have resulted in a different set of predictors (including entirely different predictor variables) being selected because the model selection process was more constrained compared to usual practice	*'I do think we probably would have tried more variations on the model structure and variables without the preregistration. After starting with standard data checks to determine appropriate model structure, I would usually run simple models to see if I could get them to work. I would have then added complexity incrementally adjusting my model as I went and troubleshooting problems as they arose. I would have adjusted the definition of the regime variable, in particular, based on ecological expectations and model outputs (if required)'* — CJ
⊕	**Adaptive Preregistration increases research transparency** Specifying modelling and analysis plans in a preregistration together with justification of choices, ensures decisions are documented and reported in greater detail. Typically, only the final models are reported	The use of interim preregistrations of the Preliminary Analysis ensured that analyses of exploratory nature were documented when they otherwise would not have been, the results of which were expressly linked to candidate model structure selection in the preregistered Main Analysis. In particular, the project lead noted that early stages of the analysis were documented when typically, decisions and analyses at these early stages, such as data cleaning and processing decisions, would not have been reported	*'We would have conducted an extensive data review in the same way as with the preregistration, but the process would not have been as well documented and that would have made it much more likely that we would have had gaps in the process and therefore more potential for errors'* — CJ

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕	**Adaptive Preregistration increases research accountability** Documenting the process helps demonstrate accountability to funders and decision makers	The project timeline included several reporting milestones and review processes. At early to middling timepoints, the modellers were able to share the preregistration document to funders to indicate how they intended to conduct the analysis, communicate rationale for particular analysis choices and demonstrate time investment	*'I was very glad to have the preregistration on hand to share or refer to for reporting milestones and review processes. The project was also long and complex with several delays, so it was nice to have a clear document outlining the process so that we could go back and check if we forgot an element of our plans'* — CJ
⊕	**Adaptive Preregistration encourages regular communication between modellers and others involved in the research** Clear communication among team members improves the chances that model outputs align with the project aims	*'The Adaptive Preregistration process (like static preregistration) requires clear communication between all people involved in collecting and modelling the data. Frequent communication provides more clarity about the purpose and scope of analyses, which may lead to model outputs being more closely aligned with project aims than would otherwise occur'* — JDY	
⊕	**Preregistration helps clarify good practice and process for less experienced researchers**	The lead and assistant modellers worked closely together during the case study, with the assistant modeller implementing part of the model development and coding of the models under the supervision of the lead modeller. The iterative nature of Adaptive Preregistration and its interim preregistrations could facilitate less experienced researchers to document their analysis choice rationale and validate this with experts or more experienced modellers before proceeding, so they could progress with confidence	*'The process helps to avoid future problems, improves transparency and quality of research practices, and is a very useful tool for less experienced researchers wanting to develop their skills. I strongly encourage everyone to go through it to understand how it works and how it could be of benefit to themselves and/or their colleagues'* — CJ
	**Develop the preregistration and analysis over a short time period** so that institutional or organisational shifts do not disrupt the research process	*'The case study presented here was developed over several years, during which time there was staff turnover and changes in project constraints (including budgets). This complicated the process'* — JDY	

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕	**Adaptive Preregistration provides a paper trail and archive of the current state of the project,** facilitating swift continuation of progress in case of unexpected loss of key personnel	*'It would have been extremely important to have the preregistration document on hand if something happened to our biometrician (e.g. sickness or other absence, which is common). If we had their code and the preregistration document, another biometrician would have been able to very quickly get everything they needed to progress things'* — CJ	

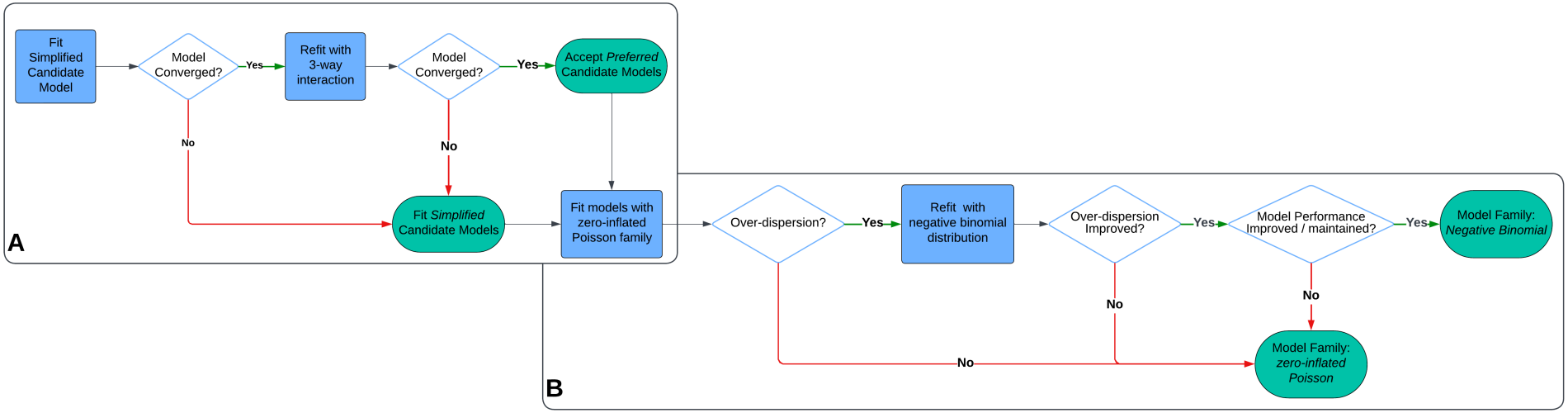


Figure 5.4: *Application of registered flexibility within the case study preregistration, for selecting model structure (a) and selecting an alternative functional form over the default zero-inflated Poisson model (b).*

5.4 Discussion

The potential benefits of preregistration are clear: it encourages careful and considered analysis, restricts researcher degrees of freedom and mitigates questionable research practices (QRPs) that can derail research. Preregistration can improve the reproducibility and reliability of scientific research (Koivisto & Mäntylä, 2024).

Existing preregistration templates are best suited to hypothesis testing research, and less suited to the style of analysis often taken by ecologists: that is, iterative modelling of complex relationships, often switching between exploratory and confirmatory practices (Alspaugh et al., 2019; Connolly et al., 2017; Prosperi et al., 2019). Although iteration makes preregistration less approachable for model-based studies, it need not preclude it entirely. Here, we show how Adaptive Preregistration can facilitate best practice via a structured, principled and flexible approach to analysis.

Despite fears that preregistration may limit creativity and flexibility (Pu et al., 2019) and lessen researchers’ engagement with data (MacEachern & Van Zandt, 2019), our case study shows how Adaptive Preregistration can improve reproducibility and transparency of process, research outputs and analytic decision-making (Liu et al., 2020). By separating the planning stage from the analysis, our case study researchers found the structured approach focused modelling and analysis activities, maintaining clear links between methodological decisions and the study aims and conceptual models while providing necessary flexibility to avoid committing to unsuitable modelling approaches (Table 5.2).

By facilitating data exploration in modelling through registered flexibility, data-subsetting and interim preregistrations, Adaptive Preregistration trades unbounded explorations for more goal-directed analyses, constraining researcher degrees of freedom and ‘fishing expeditions’. In the absence of Adaptive Preregistration, the case study modellers felt they would have started with a much larger decision-space, trialling more variations on model structure and predictor variable operationalisations along the way, risking overfitting (Lewis et al., 2023).

While Adaptive Preregistration may increase overfitting risk compared to static preregistration due to its greater flexibility, it represents a substantial improvement over current practice (no preregistration) and includes specific safeguards against the QRPs that lead to overfitting. Through registered flexibility, Adaptive Preregistration prevents opportunistic

QRPs, like ‘model fishing’, by making data-dependent decisions systematic rather than opportunistic, following transparent and prespecified criteria for modelling choices rather than optimising for desired outcomes (Chapter 4). Importantly, by formally incorporating exploratory activities into the preregistration process, Adaptive Preregistration exposes preliminary analyses that may involve QRPs but remain hidden in static preregistration approaches, where such exploration occurs before — and outside of — the formal preregistration. Our findings also show that Adaptive Preregistration increases both analytic and model transparency (Bodner et al., 2020; Lupia & Elman, 2014) by ensuring that modelling choices, steps, assumptions and expectations about outputs are articulated in greater detail than would typically occur (Schmolke et al., 2010). Documenting intermediate models, analyses and results exposes the influence of exploratory and data-contingent analyses on modelling choices. In our case study, the use of explicit prompts for choosing between and transparently justifying alternative model functional forms illuminated the analytic garden of forking paths, which too often remains opaque in ecological modelling (Fitzpatrick et al., 2021; Schmolke et al., 2010)

Modelling is considered an ‘art form’ as well as a science (Smaldino, 2020). Unpacking ingrained tacit knowledge may take time when writing preregistrations, especially the first time around, but there is clear value to capturing this expertise for future researchers working on the same or similar problems. It took several attempts for our case study modellers to fully articulate registered flexibility when selecting model structure and functional form, perhaps because these procedures were second-nature in the modellers’ practice, but by directing users to specific modelling tasks, preregistration prompted modellers to explicitly recognise, document and justify tacit modelling choices.

5.4.1 When is Adaptive Preregistration useful?

We emphasise that Adaptive Preregistration is not intended to replace static preregistration for research that can be adequately served by existing static approaches. Rather, our case study represents a particularly challenging scenario for static preregistration: the modelling team was unfamiliar with the vegetation response data structure, had limited prior information about appropriate model forms for this system and faced institutional constraints requiring hypothesis-driven conclusions rather than purely open-ended exploratory analysis. These conditions — common in applied ecological research — made it impossible to specify a complete analysis plan a priori without risking commitment to unsuitable

approaches. The case study exemplifies both information and procedural dependencies that Adaptive Preregistration is designed to address (Section 5.3.2): Preliminary Analysis results informed model structure choices (information dependency), and certain model types (zero-inflated Poisson, negative binomial) only became relevant after Preliminary Analysis revealed specific data properties (procedural dependency).

Whereas existing approaches to preregistration work well for familiar analytical contexts, we see Adaptive Preregistration as most beneficial when research involves these types of legitimate sequential decision-making and/or high data-model uncertainty, where researchers cannot reasonably predict whether intended models will be supported by unfamiliar or complex datasets. In such cases, static preregistration would force researchers into speculative commitments or require extensive post-hoc explanations of deviations.

5.4.2 Difficulties encountered using Adaptive Preregistration in our case study

Our case study researchers reported that the most challenging aspects of Adaptive Preregistration involved anticipating analysis decisions in advance, and changes to model management and timelines. Registering flexibility requires careful thought and time to adequately capture analytic decision-points and describe the process for resolving those decisions. In our case study, while the decision heuristic for choosing model functional form was accurately preregistered, the process for determining model structure was underspecified in the Main Analysis preregistration. This occurred because the preregistration did not specify that the heuristic should only apply to some models and not others (i.e. flow event models and not flow regime models, Figure 5.3).

The case study researchers reflected that anticipating potential analysis scenarios and decision pathways was complicated by large uncertainties about the underlying system dynamics in a complex and patchy dataset. The preregistration, while detailed, did not fully capture the modellers' intended and implemented model selection strategy, where after fitting the set of simpler models identified from the Preliminary Analysis (preregistration V2, Table 5.1), the modellers iteratively trialled different combinations of model terms to generate models closer to the more complex 'maximal' models identified in the initial preregistration (Table 5.1). The preregistered specification did not explain which combinations of model terms would be trialled, and in what order, rendering aspects of the model selection process unregistered. Project milestones with tight project turnaround

times left limited time to iteratively review analysis specification decisions, resulting in some components of the modelling process being omitted from the preregistration. In an ideal world, following our Adaptive Preregistration user guide more closely may have avoided under-specificity in the case study preregistration, but our case study faced the same resource constraints typical of the real world, for example, costs required to upskill some members of the research team with git and GitHub.

Our second aim for the template and Adaptive Preregistration methodology was to facilitate 'preregistration checking'—the process of verifying that the analyses reported in a study matched what was preregistered (Pu et al., 2019). We approached preregistration checking through the Adaptive Preregistration methodology itself, specifically, through the registered flexibility and interim preregistrations mechanisms, together with our version-control protocol. These facilitated comparison of the data-dependent modelling pathways realised in the course of the study against the full set of decision alternatives at each decision-point, providing explicit rationale for any data-dependent preregistered analysis. By version-controlling the preregistration document, the genesis of the preregistration from one version of the next is made explicit, for example, GitHub's 'diff view' illustrates exactly how and where a document has changed between versions (Figure 5.1, Appendix D.3, Figure 1, p.102). Also, the use of GitHub issues in our protocol tracks the findings of each analysis and links them to the corresponding preregistration item. By viewing each preregistration item's GitHub issue, the implemented analyses and findings, together with any subsequent analyses, is clear to anyone checking the preregistration. Due to limitations in our own case study application (discussed above), the analysis paths and rationale were only transparent for some preregistration items (e.g. preregistered data cleaning and processing tasks, Figure D.11, Section D.6) or only partially transparent, with discussion of some changes to the analysis plan being recorded transparently (e.g. Figure D.12, Section D.6), but not explicitly linked to changes to the modelling and analysis code itself because changes to the modelling code were not tagged under the relevant GitHub issue. Consequently, we only partially achieved our aim of facilitating preregistration checking.

These aspects combined may result in researchers feeling like Adaptive Preregistration is slowing the research process because preregistration front-loads decision-making and planning in the project timeline, delaying analysis and implementation of the preregistration (Evans et al., 2023), which may be incongruent with existing institutional approaches to the governance, administration and operational support of modelling (i.e. 'model management,'

see T. Arnold et al., 2020, p. 2). In hindsight, the case study preregistration process may have been more efficient and better specified if modellers and/or biostatisticians were brought earlier into the project's planning process to allow sufficient time to review, record and better connect analysis decisions to project aims. The case study was a particularly complex analysis with which to apply and test Adaptive Preregistration because of the data constraints and associated model structural uncertainty. For simpler analyses, Adaptive Preregistration would likely be less difficult once the process is broadly understood. We recognise that learning and implementing a new approach is unappealing at first, and although Adaptive Preregistration may save time later, it may require a rearrangement of project timelines and resources. Based on our reflections on the case study, we outline some advice for researchers wishing to implement Adaptive Preregistration in Box 1.

Box 1: Tips for doing Adaptive Preregistration

Adaptive Preregistration should become easier and more efficient as researchers gain experience and familiarity with preregistration. When implementing Adaptive Preregistration for the first time, we recommend the following:

1. Start with simple descriptions or dot points in preregistration responses, building detail and complexity in analysis specification incrementally, as well as thorough and ongoing collaborative review.

2. Refer to existing preregistrations to guide researchers in completing their own. Although there are few examples in ecology now, with future uptake and an increasing number of journals offering the Registered Reports publication pathway (e.g. Conservation Biology, Biological Conservation, Journal of Ecology and Nature Ecology and Evolution), readers should expect an expanding pool of preregistrations to draw on. Existing registration repositories, such as OSF Registries (osf.io/registries/discover), can also be searched.

3. Developing refined templates for specific applications (e.g. prediction vs. inference, or specialised complex methods and study designs) may be necessary for adequately capturing analytical decisions. In contrast, with time and community uptake, a set of decision-points common across modelling methods and applications may naturally evolve over time through practice. Our template offers a starting point from which researchers can use, modify and add to as they see fit.

4. For some researchers, using version-control for transparently documenting the Adaptive Preregistration process may be too challenging. In such cases, or for simpler analyses, sequential analysis plans could be preregistered separately for the same study with metadata linking the chain of preregistrations without tracking associated analysis files. The OSF currently allows for preregistration updating after submission (but not updating of project files, help.osf.io/article/330-welcome-to-registrations#Update-a-Registration-Z6Qdx).

5. Any first attempt to implement Adaptive Preregistration is unlikely to work perfectly, there will be mistakes and details that are omitted. Being upfront about this in study reporting is still better than avoiding preregistration entirely. See Lakens (2024, fig. 1) and Willroth & Atherton (2024, Table 1) for guidance on reporting preregistration deviations.

5.4.3 Future Work

While we believe we have partly addressed the problem of preregistering model-based research with the Adaptive Preregistration protocol, further work remains, particularly around determining the appropriate resolution of the preregistration template and developing cyber-infrastructure to effectively and transparently handle registering and reporting flexible analyses.

What is the right resolution for a modelling preregistration template?

We aimed to design a preregistration template that was ‘parsimonious’, in that (1) template items should be specific and exhaustive enough to adequately constrain researcher degrees of freedom (Wicherts et al., 2016) so as to (2) facilitate transparent documentation of the modelling paths and decisions conducted during research, while (3) not requiring a prohibitive level of detail. Increasing the completeness and resolution of the template in capturing decisions throughout the modelling process may improve the ability of preregistration to restrict researcher degrees of freedom and transparently report results; however, it also makes the preregistration more challenging to complete. Refining the template during user testing greatly reduced the perceived difficulty of preregistration. However, our

case study researchers conceded that some modellers may still consider the final template to be too cumbersome to invest the time in preregistration.

The writing and reporting phase of the case study also highlighted the need for ongoing refinement of the template content and structure in pursuit of our second use-case for the Adaptive Preregistration template: ‘preregistration checking’. Registered flexibility for determining alternative model functional forms was spread over multiple subsections within the case study’s preregistration template, hampering our ability to conduct preregistration checking post-analysis (Table 5.2). We simplified the template during testing to ameliorate this issue, but it might be further improved by building modularity into the template and preregistration platform when preregistering multiple or complex models, such as in our case study.

We encourage researchers to preregister their research using our template and Adaptive Preregistration protocol flexibly, adopting elements that work and experimenting with alternative implementations that work better. New preregistration templates for expanding application of preregistration-specific modelling approaches and methods could be developed by integrating our template with existing reporting checklists and guidelines, such as ODMAP for reporting Species Distribution Models (Fitzpatrick et al., 2021). Alternately, researchers may identify a reduced version of our template with a minimum set of items that balances tradeoffs between an exhaustive and specific template and the resource burden of prespecifying the analysis.

Encouraging implementation and uptake of Adaptive Preregistration

Fostering wider adoption of (adaptive) preregistration will require coordinated action across the research ecosystem (O’Dea, Parker, et al., 2021): researchers adopting and promoting preregistration as best practice, funders recognising and seeking out these quality standards, and journals encouraging uptake through publication preferences and recognition systems. Journals can support uptake by expanding Registered Reports programmes to accommodate Adaptive Preregistration, developing editorial guidelines for evaluating preregistered model-based studies and training reviewers to assess registered flexible analyses. Professional societies, like the Society for Open, Reliable, and Transparent Evolutionary Biology (SORTEE, sortee.org), have an important role in promoting best practices through training workshops and recognition schemes that celebrate transparent research practices. Engaging funders requires demonstrating that preregistration improves study quality, efficiency and reputation.

Researchers can proactively market their preregistration practices as a quality standard when competing for funding, positioning themselves as the ‘gold standard’ for rigorous research.

The development of appropriate registries for Adaptive Preregistration is a critical barrier to address. Existing registries, such as the Open Science Framework (OSF, help.osf.io/article/330-welcome-to-registrations) , OSF Registries (osf.io/registries), As Predicted (aspredicted.org) and EcoEvoArXiv (ecoevorxiv.org), cannot accommodate Adaptive Preregistration, which requires dedicated instruments for capturing flexible analyses, reporting their outcomes and linking them to downstream decisions in subsequent preregistrations. To circumvent the inability to integrate Adaptive Preregistration into existing registries, we described and demonstrated in our case study a protocol for implementing Adaptive Preregistration and facilitating preregistration checking using git and GitHub. However, GitHub does not meet the minimum requirements for FAIR (Findable, Accessible, Interoperable and Reusable) data archiving (Pick et al., 2025), nor for preregistration (Haroz, 2022). Existing registries must be modified or new registries created to enable Adaptive Preregistration. Future registries could further streamline Adaptive Preregistration by displaying modular template content, shown to the user conditionally based on their preregistration responses, such that the preregistration ‘expands’ incrementally as items are completed. Publishers are well positioned to develop or commission (adaptive) preregistration registries that integrate with their submission systems, creating transparent linkages between the preregistration and final paper.

As open science practices are becoming increasingly embedded in the publication process, for example, reproducibility checklists at Conservation Biology and data- and code-checking in the peer-review workflow at Ecology Letters (Thrall et al., 2023), preregistration should further streamline the publication process by enhancing the transparency and rigour of the study. Beyond preregistration infrastructure, publishers could support this transition by developing standardised guidelines for reporting registrations in resulting manuscripts and providing clear editorial policies on how preregistered studies (that are not Registered Reports) should be handled during peer review. Such guidelines could address how to clearly reference preregistration items in manuscripts, report deviations transparently and present results from registered flexible analyses. For example, manuscripts should ideally reference the corresponding preregistration item number in the methods and results sections,

explicitly acknowledging when and why any deviations occurred (e.g. see Chapter 2, where deviations were reported with text-boxes).

5.4.4 Conclusions

Although challenges remain in preregistering ecological modelling, it is important to remember the costs and risks of not doing so. Ecological modelling has long been plagued by poor transparency. Incomplete reporting of the modelling process risks models being used in inappropriate applications or decisions and may mask poor model design or serious flaws in the model, resulting in adverse or irreversible outcomes when informing decision-making. Moreover, a lack of transparency in describing ecological models and reporting the modelling process provides significant opportunity for undisclosed researcher degrees of freedom, and hence, the possibility of QRPs, with researchers routinely executing alternative analyses and selectively reporting them when they do. Reproducible, transparent and reliable models are essential for sound conservation decision-making, and Adaptive Preregistration can improve documentation of modelling decisions, which is helpful for remembering and explaining the study’s method at the time of report-writing, to aid others in understanding and replicating model-based research and to communicate to funders and stakeholders.

We provide a template and methodology for Adaptive Preregistration that we hope will extend the benefits of preregistration to model-based research in a way that strengthens rather than inhibits that research. Although we focus on ecological applications in this paper, many elements of the template are relevant across fields, and the open-source template can be readily adapted to new fields and purposes.

6 General Discussion

In this thesis, I set out to investigate strategies for improving reproducibility and transparency in ecological modelling. I have focused especially on understanding the nature and consequences of analytic flexibility or researcher degrees of freedom. The nature of analytic flexibility and the role that transparency can play in mitigating it warrant deep investigation in part because these aspects of open science have become highly politicised (Zagrodzka et al., 2024). Scientific findings that contradict political or economic interests face increasing scrutiny, and strategies for dismissing unwelcome evidence now sometimes employ strong open science rhetoric (Nosek, 2025).

The rise of scientific populism and the co-option of open science rhetoric to reject scientific evidence is not restricted to any one country (Center for Open Science, 2025; Cologna et al., 2025; Koziol, 2025). In Australia too, the 'reproducibility crisis' has been weaponised by those seeking to undermine the epistemic authority of science (Bush et al., 2019). Nosek (2025) offers us a rallying perspective to combat this double bind; proffering that embracing open science, rather than retreating from it, will strengthen public trust by revealing science's slow and arduous process through which knowledge is incrementally contested and built. My work is a contribution made in that spirit, and with the purpose of maintaining public trust in the science underpinning environmental decisions. This is especially critical for ecological modelling, which plays an increasingly central role in guiding conservation strategies, informing environmental policy, and supporting decision-making in the face of global environmental change.

Concerns about reproducibility are not merely scientific concerns but are critical requirements of effective conservation science (Powers & Hampton, 2018). Transparent and reliable ecological models are critical for building a robust evidence base informing management

(Dicks et al., 2014; Gardner et al., 2018; Law et al., 2017; Sutherland & Wordley, 2017). Poor modelling practices can lead to inappropriate inferences or inaccurate predictions, which may lead to sub-optimal allocation of resources or damaging outcomes (Grimm et al., 2014; Morrison et al., 2016; Schmolke et al., 2010; E. G. Simmonds et al., 2024; Wood et al., 2020).

Before a final discussion of the contributions this work makes to the field, I will briefly review the empirical chapters, drawing out the primary lesson in each. In Chapter 1, I lay the foundation for this thesis by examining how threats to reproducibility manifest in ecological research, documenting how the combination of extensive researcher degrees of freedom and a culture of opaque model reporting has created the conditions for a 'reproducibility crisis' – in particular, questionable research practices (QRPs). I examined existing open science efforts in ecology, which have primarily focused on computational reproducibility through data and code sharing, thereby highlighting a fundamental gap in addressing transparency and reproducibility challenges within model-based research, specifically; existing frameworks for understanding and mitigating these threats to reproducibility and transparency are poorly suited to ecological modelling contexts. Through this analysis, I established the central tension of the thesis. That is, how to constrain researcher degrees of freedom to ensure reproducible and transparent results, while preserving the methodological flexibility required by complex ecological modelling?

My first empirical chapter, Chapter 2, demonstrated both the presence and potential of researcher degrees of freedom to generate heterogeneous results. In a large-scale 'many analysts' study, I quantitatively illustrated the garden of forking paths for two separate research questions and datasets, highlighting opportunities for QRPs. This work provided the first empirical quantification of analytical heterogeneity in ecological research, demonstrating that analytic flexibility can substantially alter inferences and conclusions even among competent researchers using the same data. Critically, this study moved beyond theoretical concerns about reproducibility to provide concrete evidence of how analytical decisions affect research outcomes in ecology. The findings challenge assumptions about the reliability of results from a single study or analysis specification, underscoring the need for i) further metaresearch to understand the scope and consequences of researcher degrees of freedom in ecological contexts, and ii) methodological reforms in ecological research practices.

My approach to conducting the research in this thesis has been guided by the principle of 'practising what I preach,' following best practices in data and code sharing, and ensuring the computational reproducibility of quantitative analyses. As such, there are non-traditional research outputs accompanying many of my chapters. I developed an R software package (Chapter 3) for conducting the Same Data, Many Analysts study in Chapter 2, in which I applied the software to create a fully reproducible analysis pipeline and manuscript. The software facilitates reproduction and verification of the analysis in Chapter 2 and allows others to re-analyse our data or replicate our study on new data.

Building on the empirical evidence from Chapter 2, in Chapter 4, I addressed the conceptual gap in understanding how QRPs manifest in ecological modelling contexts, specifically. I extended the formalism developed by Gelman & Loken (2013) describing the 'garden of forking paths' and applied it to the case of ecological modelling, establishing that QRPs may threaten the credibility and reliability of model-based research, just as they do in hypothesis-testing research. However, I argued that ecological modelling is especially at risk of QRPs because of its culture of poor model reporting practices that yield a severe lack of transparency. This chapter overcame significant conceptual barriers by moving beyond hypothesis-testing frameworks to address the incompatibility of existing QRP definitions with model-based research and by distinguishing systematic from opportunistic data-dependent decision-making in iterative modelling processes. By resolving these conceptual issues, I provided the first definition of QRPs applicable to ecological modelling, described potential mechanisms by which QRPs emerge, and identified modelling-specific QRPs at different points in the modelling process. This work established that ecological modelling has its own distinct set of QRPs that modellers should be aware of, providing the conceptual foundation necessary for developing solutions tailored to ecological modelling. With the aid of our map of QRPs, modellers are equipped to avoid common or high-risk QRPs in their work, while readers can spot potential QRPs in published model-based studies.

As established in other disciplines, awareness of QRPs is insufficient to stop them from occurring when conducting research. Preregistration is a commonly proposed open science method for preventing QRPs. In Chapter 5, I investigated the feasibility and potential for preregistration to mitigate the risk of QRPs and improve the transparency and reproducibility of ecological modelling. I unpacked modellers' concerns about the applicability and tractability of preregistration in ecological modelling to identify specific barriers to preregistration that might be addressed through technical means. In this chapter, I develop,

apply, and evaluate *Adaptive Preregistration*, a structured, principled, and flexible methodology for preregistration that constrains researcher degrees of freedom without curtailing the data-dependent nature of analytic decision-making in ecological modelling. Building directly on the 'systematic versus opportunistic' distinction for defining QRPs established in Chapter 4, the Adaptive Preregistration template in Chapter 5 provides structured methods to document when data-dependent decisions are defensible, thereby mitigating QRPs while preserving the adaptive nature of ecological modelling. The application of the template and Adaptive Preregistration methodology to the environmental flows modelling case study is an important proof of concept for demonstrating its practical utility.

6.1 Recommendations for Research Practice

The work presented in this thesis leads to three primary and interlinked recommendations for research practice, which will be most effective when adopted in tandem, rather than in isolation. These recommendations tackle interconnected challenges in ecological modelling: first, reducing ambiguity in research questions that can mask legitimate analytical differences and create opportunities for researcher degrees of freedom to be exploited; second, enhancing model transparency to enable identification and prevention of QRPs; and third, developing tools and infrastructure to make Adaptive Preregistration practical for iterative modelling processes. As I present each, I will discuss implementation challenges and propose solutions to advance the reproducibility and reliability of ecological modelling, as well as suggesting future work that might inform these changes.

6.1.1 Improve Precision and Clarity of Research Question Framing

Like Rohrer et al. (2025), I argue that the value of many-analyst and multiverse studies lies not in their capacity to yield a specific measure of the reliability or credibility of studies on a particular research question or field, but rather, to yield insights about current practice, including limitations. In line with Borger & Ramesh (2025), I propose that our reaction to many-analyst studies should shift from shock at observing variation in apparently statistically valid and plausible analyses to understanding *why* differences in results are observed, that is, identifying the conceptual differences among studies, if any.

As acknowledged in Chapter 2, some of the variation in approaches to analysis may have stemmed from broad research questions posed to analysts. While research questions posed

in this study were designed to match the level of specificity in the published literature, a common criticism of many-analyst studies is that the imprecision in research questions gives space for researchers to explore different ecological questions (Knief & Forstmeier, 2025). In their targeted multiverse analysis of our blue tit dataset, Knief & Forstmeier (2025) argued that truly arbitrary analytical decisions yielded similar effect sizes, while analyses yielding heterogeneous effect sizes effectively operationalised different research questions. Yet, Rohrer et al. (2025) found that, even when research questions are reformulated to be less ambiguous, high levels of heterogeneity remain in analysts' results, with estimated effect sizes and signs still varying across analyses. Nonetheless, it is clear that the precision of the formulated research question, not only in many-analyst studies, but in published literature in general, warrants further consideration.

These findings perhaps underscore the importance of clearly articulated research questions in general. Vague research questions - those that suggest a pattern without specifying causal processes or conditions under which that pattern holds - cannot reliably constrain the model specification space, leading to different model specifications that capture different data-generating processes. Where vague questions fail to specify an estimand, researchers may choose different response variables to represent the construct of interest. In Chapter 2, for instance, analysts chose different size classes of seedlings as their response variable that may reflect different stages of the recruitment process, from germination, early establishment, survival to sapling stages. Failing to encode different causal assumptions in the model makes it difficult to compare effect sizes. Also, when the research question omits the underlying process, researchers lack a principled basis for selecting predictor variables and may resort to including all variables available in the dataset, regardless of ecological plausibility or causality ('causal salad' approaches, McElreath (2020), see Section 4.6.3).

The problem of vague research questions parallels with the emerging *theory crisis* in psychology, in which studies fail to replicate because researchers test vague hypotheses due to poorly described verbal theories (Nakagawa et al., 2025; Yarkoni, 2022). The intersecting problems of variable selection that ignores ecological relevance (Fourcade et al., 2018), and of vaguely specified hypotheses, are not new in ecology (e.g., Yanco, 2020). However, the role of vague theories in contributing to these problems, and the potential impacts on replicability, have only recently been articulated (Nakagawa et al., 2025). One proposed method for distinguishing when variation in results is driven by arbitrary analytic flexibility

versus different ecological interpretations of research questions is by explicitly communicating conceptual models, including Directed Acyclic Graphs (DAGS, Borger & Ramesh, 2025; Franks et al., 2025; Knief & Forstmeier, 2025). By forcing modellers, decision-makers and practitioners alike to fully describe, explain and justify the inclusion of parameters, ecological processes and causal assumptions, the process of communicating underlying mental models helps to move discussion towards conceptual differences that motivate differences in modelling choices (Borger & Ramesh, 2025). Articulating conceptual models may also remove temptation and opportunities for HARKing and other QRPs, especially when they are articulated at the outset of any new modelling problem. As I proposed in Chapter 5, explicitly describing conceptual models is a critical step in preregistering model-based research. However, transparent communication of conceptual models is not enough alone to improve transparency and reproducibility.

6.1.2 Improve Model Transparency to Reduce the Risk of QRPs, Facilitate Model Reuse and Synthesis

Although the QRP roadmap (Chapter 4) constitutes a first step towards raising awareness among ecological modellers of QRPs throughout the modelling process, the impact of this work is potentially hampered by a poor culture of model-reporting in ecological modelling (Chapter 1). Simply put, ecological models are so often incompletely and opaquely reported that it is impossible to spot the presence of QRPs – if modelling decisions remain hidden, readers cannot identify whether systematic or opportunistic decision-making has occurred, or whether results have been selectively reported. Yet poor transparency is a routine and normalised practice in ecological modelling, so a lack of transparency alone is not necessarily a red flag for QRPs.

Improving research transparency is therefore a critical prerequisite to identifying QRPs in research, as well as preventing their occurrence. Open science efforts in ecological modelling should focus on developing reporting and peer-review checklists, and consider implementing reporting tools, like 'model cards.' Model cards are an alternative to reporting checklists developed for use within the Machine Learning and AI research communities. They transparently and succinctly document the model, its limitations and intended scope for application, the model development methods as well as a summary of evaluation results (Garbin & Marques, 2022; Mitchell et al., 2019). Reporting checklists and model cards could borrow from the structure and content of the Adaptive Preregistration template from

this thesis, or else be tailored to specific modelling methods. Metaresearch examining the relative extent and impact of QRPs in ecological modelling should inform the content of new transparency checklists or model cards, while policy will be required to promote their use by journals and study authors.

Adaptive Preregistration documents may serve as a source of structured data on how ecological modelling is actually conducted. Coupled with a suitable template, Adaptive Preregistration can generate rich, structured records of the modelling process that go beyond what is typically reported in published methods sections. In particular, registered flexibility reveals the iterative and adaptive nature of modelling in practice, while providing documented decision rationales that explicitly link modelling choices to study objectives. Although normative and prescriptive guidance about 'modelling process' abound in the ecological modelling literature, Lahtinen et al. (2017) identify that the realised sequence of decisions or 'modelling paths' may look very different in practice to the 'best-practice' guidance the researchers set out to follow. Machine-readable preregistrations (e.g., Lakens & DeBruine, 2021; Van Eijk et al., 2023) could support metaresearch by enabling systematic analysis comparing how modelling is conducted in practice. This data could be compared against the reported analyses and results to identify deviations from the preregistered analysis, or potential QRPs like selective-reporting. Preregistration registries could provide a structured database for AI-based tools to learn from existing preregistrations, guiding researchers through planning, conducting and reporting complex modelling workflows, particularly those with limited formal training.

The risks of such tools also warrant consideration: systems trained on accumulated preregistrations risk privileging established methodological conventions at the expense of legitimate diversity in modelling approaches, embedding the biases of current practice rather than challenging them. Secondly, mechanisms for consenting for inclusion in the training database and providing appropriate attribution are essential for ethical reuse of published preregistrations. Open science and metaresearch have a role to play in both ensuring that any such training corpus is representative of the full breadth of ecological modelling contexts and purposes, and in empirically evaluating whether AI-assisted guidance genuinely improves practice or merely standardises it.

To aid in prioritising which components of modelling methods and results should be included in reporting checklists, future metaresearch could include surveys of the ecological

modelling community, like that of Fraser et al. (2018). These surveys would reveal which QRPs are routinely executed, and which are of most concern to the modelling community. Further work to characterise the severity or impact of QRPs on biasing model-based research might include in-silico simulation experiments in which models are systematically altered to simulate the impacts of QRPs on the model and analysis outcomes. Alternatively, in-silico experiments or specification curve analyses could investigate the effect of QRPs on biasing the overall literature to identify practices of critical concern (e.g., Desbureaux, 2021; McElreath & Smaldino, 2015; Nissen et al., 2016; Simonsohn et al., 2020).

Self-report surveys of modellers about QRP engagement could ask about practices in applied contexts (e.g., reports to government), not just about academic publications. The QRP framework in Chapter 4 and the Adaptive Preregistration methodology in Chapter 5 both treat publication success as the primary driver of QRPs, but reflections from the case study modellers suggest that qualitatively different pressures operate at the research–practice interface that may drive distinct forms of opportunistic decision-making. Distinct pressures included compressed timelines, institutional reporting milestones, and funder expectations. Characterising these context-specific incentive structures, and their implications for which QRPs are most likely in decision-support settings, is therefore a natural extension of the survey agenda, using the QRP classification developed in this thesis adapted to capture decision-relevant outputs and practitioner-specific pressures.

Given there is a practical limit to research transparency – it is simply intractable to report and unreasonable to read the entire modelling process in all its minutiae – other solutions beyond improving transparency are necessary to mitigate or prevent QRPs. In this thesis I have demonstrated how Adaptive Preregistration can be used to prevent QRPs in model-based research, but preregistration can be coupled with other techniques, such as 'data blinding' or 'data scrambling' (Baldwin et al., 2022; Traylor, 2024), to minimise the potential for outcome-dependent decisions. The metaresearch agenda should consider quantitatively evaluating the effectiveness of these reforms in mitigating QRPs and reducing bias in results.

6.1.3 Refine Adaptive Preregistration Infrastructure & Templates to Facilitate Preregistration Uptake

The Adaptive Preregistration case study highlighted the tension between the ideals of a good preregistration template and practical feasibility of conducting preregistration. This

tension is likely to be particularly salient in applied modelling contexts, where there is a need to act rapidly with limited resources and fast turnaround times on modelling and analysis. The case study preregistration suffered from underspecificity in describing the model selection procedure, and omission of some candidate models and analysis steps. In addition, only some data-dependent decisions were transparently documented using the git and GitHub protocol, and not all elements of the protocol were followed due to insufficient time for upskilling some team members within tight project turnaround times. My protocol for implementing Adaptive Preregistration with git and GitHub therefore was only partly successful in facilitating 'preregistration checking' (that is, comparing the preregistered versus reported analysis). These shortcomings highlight the need to develop cyberinfrastructure that better accommodates interim preregistrations and registered flexibility, and to focus on recording only critical analytical decisions to avoid overburdening researchers.

Beyond its intended function of improving research transparency and mitigating QRPs, Adaptive Preregistration may also serve an educational function. By requiring modellers to articulate their analytical rationale before analysis begins, like including anticipated decision points and responses to predictable modelling problems, the methodology naturally scaffolds structured reflection on modelling practice. Secondly, the preregistration template is intended to be somewhat prescriptive and normative, in that its structure and content is informed by best-practice guidance on model development and reporting. Evidence for a pedagogic function emerged in the Chapter 5 case study: the project lead reflected that the process is "a very useful tool for less experienced researchers wanting to develop their skills," noting that the iterative interim preregistration structure would encourage less experienced modellers to document and validate their analytical reasoning with more experienced colleagues before proceeding (Table 5.2). The lead modeller similarly noted that the front-loading and explicit prompting for decision rationale imposed by Adaptive Preregistration improved clarity about the scope and purpose of analyses by encouraging communication across the team. Future work evaluating the promise and the limits of Adaptive Preregistration as training and/or supervisory tool (Blincoe & Buchert, 2020) would be a valuable contribution.

I have demonstrated that preregistration of model-based research is possible but comes with challenges. Realistically, widespread adoption of Adaptive Preregistration requires further refinement of the template to make it easier, especially in applied decision-making settings. The template I developed is especially cumbersome because it is skewed towards

completeness and exhaustiveness. Further research is needed to address the knowledge-action gap and develop a 'rapid template' for decision-support contexts (Roche et al., 2021). While a 'reduced template' for use in academic contexts is needed to lower the burden of preregistration on already strained researchers (O'Dea, Parker, et al., 2021). Both projects will require explicit decisions about the trade-off between restricting researcher degrees of freedom and complete research transparency. Moreover, the modelling community may want to develop distinct templates for specialised methods or modelling problems, such as a template for Species Distribution Modelling, or studies using Bayesian hierarchical modelling.

The Adaptive Preregistration template included elements of best-practice guidance for decision-support contexts (e.g. articulation of modelling objectives, problem context, and explicit links between modelling choices and those objectives, Appendix D) and it was tested on a decision-support modelling case study. It primarily captured decisions about model construction and evaluation, and it could potentially be expanded to include evaluation of management options and tradeoffs. This would perhaps improve its uptake for Structured Decision Making, Adaptive Management, and value-of-information analyses.

Priorities for future metaresearch may include conducting a systematic review of how objectives, decision alternatives, and decision criteria are currently documented in published applied ecological modelling studies. Content analysis could characterise how frequently and completely the chain of reported ecological model outputs to recommended actions is reported, and whether the criteria used to translate model outputs into decisions are disclosed. This work would characterise the gap between current practice and what transparent decision-support preregistration would require, paralleling emerging transparency audits of reporting completeness in ecological modelling more broadly.

6.2 Contextualising the Work in this Thesis

At its core, improving the reproducibility and credibility of scientific research is a complex cultural change problem. The open science movement has generated a suite of new tools, methods, and processes for improving the reliability of research. Yet making tools available – solving the technical problem – is only part of the challenge. Getting people to use the tools is a more complex socio-cultural challenge. The research landscape in ecology, and in many other sciences, is pluralistic and decentralised, with community norms and

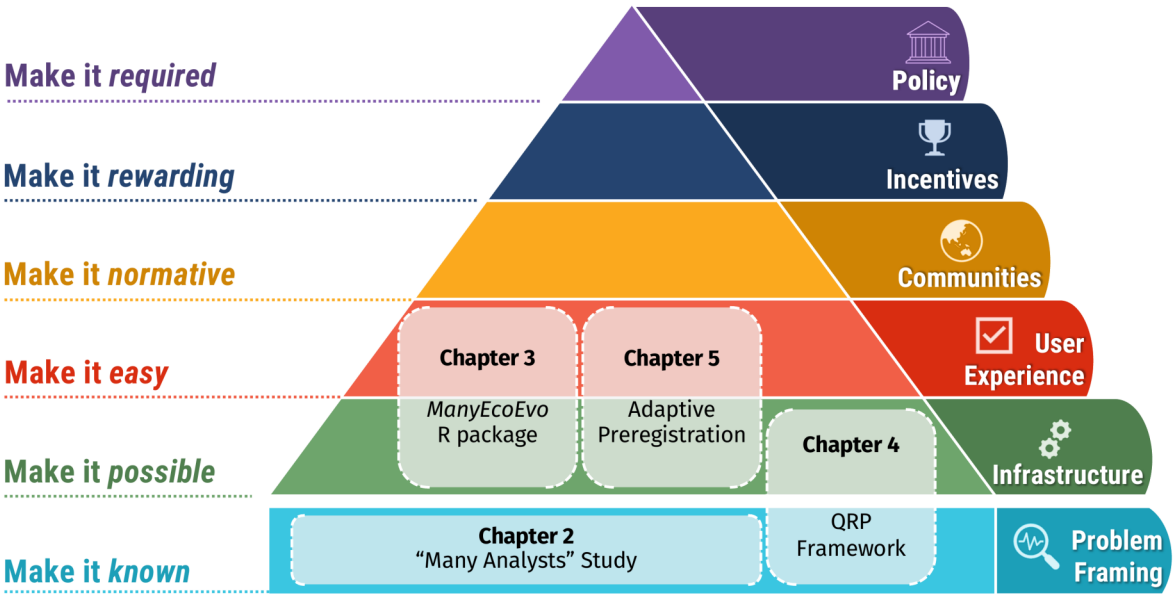


Figure 6.1: *Cuevas Shaw's (2022) strategy for change required to foster a culture of open, transparent, and reproducible science, adapted to include a 'Problem Framing' foundation that precedes the cultural change process and incorporates contributions from this thesis.*

incentive structures that resist change and maintain the status quo (O'Dea, Parker, et al., 2021). Coordinated action among all stakeholders including researchers, institutions, funding bodies, and publishers is necessary. To understand how my contributions fit within broader efforts to improve scientific practice, it is helpful to position them within established frameworks for cultural change. The Center for Open Science works with a model of cultural change illustrated in Figure 6.1.

My research sits at the foundational levels of the pyramid of change. Chapter 3 and Chapter 5 sit across *infrastructure* and *ease*, making open science both possible and easy for researchers with software and templates. In Chapter 5, I *make preregistration in ecological modelling possible* through demonstration of its application to a real applied ecological modelling problem. The accompanying preregistration template, Adaptive Preregistration protocol and user guide refined through our case study *make it easy* by providing a set of tools to implement preregistration so as to "minimise friction" (Cuevas Shaw et al., 2022), thereby encouraging adoption in the community. Chapter 4 provides *infrastructure* in the form of a map of QRPs throughout the modelling process, which makes it possible for modellers to be aware of the risk of QRPs in their research and readers of model-based research to spot potential QRPs in published modelling studies.

As Kambouris et al. (2024) and other attempts at computationally reproducing studies in ecology show, simply sharing data and code does not guarantee computational reproducibility. The combination of Chapter 2's fully reproducible quarto manuscript, and my R package *ManyEcoEvo* (Chapter 3) make it easy to reproduce, verify and re-analyse the 'Same Data, Many Analysts' data, and to replicate the study on new data. Indeed, the re-analysis of Chapter 2 by Knief & Forstmeier (2025) only several months after publication highlights my success in this pursuit.

The work in Same Data, Many Analysts study (Chapter 2), and elements of Chapter 4 sit outside the pyramid. I have illustrated this at level below or before the cultural change work begins, as 'Problem Framing' with the principle 'make it known.' I conceptualise the work in this way to emphasise that clarity around the problem itself determines how well messages transmit through the community and ensures that tools, infrastructure and policy successfully and adequately address the target problem. The work in both Chapter 2 and Chapter 4 make it known that researcher degrees of freedom threaten the reproducibility of ecological modelling research, establishing the role for preregistration in mitigating opportunities for QRPs that stem from the contingency of results on analytic flexibility. I do this acknowledging that making problems known is not enough to facilitate change in communities of practice, as is evidenced by the Good Modelling Practice's (GMP) long-standing calls to improve ecological modelling practice (Jakeman et al., 2024; Micheletti et al., 2024). The growing institutional traction of open-science policy, such as data and code sharing mandates, registered reports, and reproducibility checklists, offers a productive vehicle for achieving the practice change the Good Modelling Practice community has long advocated. By drawing these intersecting concerns into a shared discourse there is meaningful potential to amplify the reach and impact of both bodies of work.

6.3 Conclusion

This thesis provides empirical evidence, conceptual frameworks, and practical tools for transforming how ecological modelling addresses reproducibility challenges, offering a path forward for the community to build more credible, transparent, and ultimately more impactful model-based research. The primary motivation for this thesis, and its contributions, is to work towards a more solid foundation for evidence-based policy and environmental management. When faced with ever-expanding socioecological challenges, we cannot

afford to base critical policy and management decisions on opaque models with questionable reliability. While more work is needed to further illuminate the path forward, this thesis provides the first steps towards more rigorous and transparent science for improving environmental decisions.

References

A checklist for our community. (2018). *Nature Ecology & Evolution*, *2*(6), 913. https://doi.org/10.1038/s41559-018-0574-7

Achter, S., Borit, M., Chattoe-Brown, E., & Siebers, P. O. (2022). RAT-RS: A reporting standard for improving the documentation of data use in agent-based modelling. *International Journal of Social Research Methodology*, *25*(4), 517–540. https://doi.org/10.1080/13645579.2022.2049511

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B., … Wagenmakers, E. J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, *10*, e72185. https://doi.org/10.7554/eLife.72185

Addison, P. F. E., Rumpff, L., Bau, S., Carey, J. M., Chee, Y. E., Jarrad, F. C., McBride, M., & Burgman, M. A. (2013). Practical solutions for making models indispensable in conservation decision-making. *Diversity and Distributions*, *19*(5-6), 490–502. https://doi.org/10.1111/ddi.12054

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Ahmed, W., Kommineni, V. K., König-Ries, B., Gaikwad, J., Gadelha, L., & Samuel, S. (2025). Evaluating the method reproducibility of deep learning models in biodiversity research. *PeerJ Computer Science*, *11*, e2618. https://doi.org/10.7717/peerj-cs.2618

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. https://doi.org/10.1890/13-1452.1

Alexandrov, G. A., Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., Hildebrandt, A., Hoffman, F., Jackisch, C., Khaiter, P., Mannina, G., Matsunaga, T., Purucker, S. T., Rivington, M., & Samaniego, L. (2011). Technical assessment and evaluation of environmental models and software: Letter to the editor. *Environmental Modelling & Software*, *26*(3), 328–336. https://doi.org/10.1016/j.envsoft.2010.08.004

Allaire, J. J., & Dervieux, C. (2024). *Quarto: R interface to 'quarto' markdown publishing system* (version 1.4) [R package]. https://CRAN.R-project.org/package=quarto

Allaire, J. J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown: Dynamic documents for r* (version 2.27) [R package].

Allison, A. E. F., Dickson, M. E., Fisher, K. T., & Thrush, S. F. (2018). Dilemmas of modelling and decision-making in environmental research. *Environmental Modelling & Software*, *99*, 147–155. https://doi.org/10.1016/j.envsoft.2017.09.015

Alspaugh, S., Zokaei, N., Liu, A., Jin, C., & Hearst, M. A. (2019). Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

Arif, S., & MacNeil, M. A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, *93*(1), e1554. https://doi.org/10.1002/ecm.1554

Arnold, J. B. (2024). *ggthemes: Extra themes, scales and geoms for "ggplot2"* (version 5.1.0) [R package]. https://jrnold.github.io/ggthemes/

Arnold, T., Guillaume, J. H. A., Lahtinen, T. J., & Vervoort, R. W. (2020). From ad-hoc modelling to strategic infrastructure: A manifesto for model management. *Environmental Modelling and Software*, *123*, 104563. https://doi.org/10.1016/j.envsoft.2019.104563

Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution*, *35*(4), 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Atkinson, J., Brudvig, L. A., Mallen-Cooper, M., Nakagawa, S., Moles, A. T., & Bonser, S. P. (2022). Terrestrial ecosystem restoration increases biodiversity and reduces its

variability, but not to reference levels: A global meta-analysis. *Ecology Letters*, *25*(7), 1725–1737. https://doi.org/10.1111/ele.14025

ATLAS.ti Scientific Software Development GmbH. (2019). *ATLAS.ti mac* (8.4.13) [Software]. https://atlasti.com/

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "many analysts, one data set" project. *Socius: Sociological Research for a Dynamic World*, *7*, 23780231211024421. https://doi.org/10.1177/23780231211024421

Ayllón, D., Railsback, S. F., Gallagher, C., Augusiak, J., Baveco, H., Berger, U., Charles, S., Martin, R., Focks, A., Galic, N., Liu, C., van Loon, E. E., Nabe-Nielsen, J., Piou, C., Polhill, J. G., Preuss, T. G., Radchuk, V., Schmolke, A., Stadnicka-Michalak, J., … Grimm, V. (2021). Keeping modelling notebooks with TRACE: Good for you and good for environmental research and management support. *Environmental Modelling & Software*, *136*, 104932. https://doi.org/10.1016/j.envsoft.2020.104932

Babel, L., Vinck, D., & Karssenberg, D. (2019). Decision-making in model construction: Unveiling habits. *Environmental Modelling & Software*, *120*. https://doi.org/10.1016/j.envsoft.2019.07.015

Badham, J., Elsawah, S., Guillaume, J. H. A., Hamilton, S. H., Hunt, R. J., Jakeman, A. J., Pierce, S. A., Snow, V. O., Babbar-Sebens, M., Fu, B., Gober, P., Hill, M. C., Iwanaga, T., Loucks, D. P., Merritt, W. S., Peckham, S. D., Richmond, A. K., Zare, F., Ames, D., & Bammer, G. (2019). Effective modeling for Integrated Water Resource Management: A guide to contextual practices by phases and steps and future opportunities. *Environmental Modelling & Software*, *116*, 40–56. https://doi.org/10.1016/j.envsoft.2019.02.013

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. T., & Wicherts, J. M. (2020). Ensuring the

quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, *22*, 153–160. https://doi.org/10.1136/ebmental-2019-300117

Baldwin, J. R., Pingault, J. B., Schoeler, T., Sallis, H. M., & Munafò, M. R. (2022). Protecting against researcher bias in secondary data analysis: Challenges and potential solutions. *European Journal of Epidemiology*, *37*(1), 1–10. https://doi.org/10.1007/s10654-021-00839-0

Barros, C., Luo, Y., Chubaty, A. M., Eddy, I. M. S., Micheletti, T., Boisvenue, C., Andison, D. W., Cumming, S. G., & McIntire, E. J. B. (2023). Empowering ecological modellers with a PERFICT workflow: Seamlessly linking data, parameterisation, prediction, validation and visualisation. *Methods in Ecology and Evolution*, *14*(1), 173–188. https://doi.org/10.1111/2041-210X.14034

Bartoń, K. (2023). *MuMIn: Multi-model inference* (version 1.47.5) [R package]. https://CRAN.R-project.org/package=MuMIn

Baselga, A., Orme, D., Villeger, S., De Bortoli, J., Leprieur, F., Logez, M., Martinez-Santalla, S., Martin-Devasa, R., Gomez-Rodriguez, C., Crujeiras, R. M., & Henriques-Silva, R. (2023). *betapart: Partitioning beta diversity into turnover and nestedness components* (version 1.6) [R package]. https://CRAN.R-project.org/package=betapart

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 48. https://doi.org/10.18637/jss.v067.i01

Beck, M. W., O'Hara, C., Stewart Lowndes, J. S., Mazor, R. D., Theroux, S., J. Gillett, D., Lane, B., & Gearheart, G. (2020). The importance of open science for biological assessment of aquatic environments. *PeerJ*, *8*, e9539. https://doi.org/10.7717/peerj.9539

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Bennett, N. J., Roth, R., Klain, S. C., Chan, K., Christie, P., Clark, D. A., Cullman, G., Curran, D., Durbin, T. J., Epstein, G., Greenberg, A., Nelson, M. P., Sandlos, J., Stedman, R.,

Teel, T. L., Thomas, R., Veríssimo, D., & Wyborn, C. (2017). Conservation social science: Understanding and integrating human dimensions to improve conservation. *Biological Conservation*, *205*(C), 93–108. https://doi.org/10.1016/j.biocon.2016.10.006

Benning, S. D., Bachrach, R. L., Smith, E. A., Freeman, A. J., & Wright, A. G. (2019). The registration continuum in clinical science: A guide toward transparent practices. *Journal of Abnormal Psychology*, *128*(6), 528. https://doi.org/10.1037/abn0000451

Bergmann, C., Frank, M. C., Gonzalez, N., Bergelson, E., Cristia, A., Ferguson, B., Kline Struhl, M., Soderstrom, M., Yurovsky, D., Byers-Heinlein, K., Panneton, R., Lew-Williams, C., Hamlin, K., Liu, L., Dal Ben, R., Altuntas, E., Baumgartner, H. A., & Zhou, G. (2025, August). *ManyBabies*. OSF. https://osf.io/rpw6d

Betts, M. G., Hadley, A. S., Frey, D. W., Frey, S. J. K., Gannon, D., Harris, S. H., Kim, H., Kormann, U. G., Leimberger, K., Moriarty, K., Northrup, J. M., Phalan, B., Rousseau, J. S., Stokely, T. D., Valente, J. J., Wolf, C., & Zárrate-Charry, D. (2021). When are hypotheses useful in ecology and evolution? *Ecology and Evolution*, *11*(11), 5762–5776. https://doi.org/10.1002/ece3.7365

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Blake, K. (2022). *NatParksPalettes: Color palettes inspired by national parks* (version 0.2.0) [R package]. https://github.com/kevinsblake/NatParksPalettes

Blincoe, S., & Buchert, S. (2020). Research Preregistration as a Teaching and Learning Tool in Undergraduate Psychology Courses. *Psychology Learning & Teaching*, *19*(1), 107–115. https://doi.org/10.1177/1475725719875844

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A quick introduction to version control with git and GitHub. *PLOS Computational Biology*, *12*(1), e1004668. https://doi.org/10.1371/journal.pcbi.1004668

Bodner, K., Fortin, M.-J., & Molnár, P. K. (2020). Making predictive modelling ART: Accurate, reliable, and transparent. *Ecosphere*, *11*(6), e03160. https://doi.org/10.1002/ecs2.3160

Bodner, K., Rauen Firkowski, C., Bennett, J. R., Brookson, C., Dietze, M., Green, S., Hughes, J., Kerr, J., Kunegel-Lion, M., Leroux, S. J., McIntire, E., Molnár, P. K., Simpkins, C., Tekwa, E., Watts, A., & Fortin, M.-J. (2021). Bridging the divide between ecological forecasts and environmental decision making. *Ecosphere*, *12*(12), e03869. https://doi.org/10.1002/ecs2.3869

Bohn, M., Schmitt, V., Sanchez-Amaro, A., Keupp, S., Hopper, L., Völter, C., Altschul, D., Fischer, J., Fichtel, C., Beran, M. J., Kano, F., Call, J., Watzek, J., Joly, M., & Hernandez-Aguilar, R. A. (2024, January). *ManyPrimates*. OSF. https://osf.io/tr6q3

Bokulich, A. (2013). Explanatory models versus predictive models: Reduced complexity modeling in geomorphology. In V. Karakostas & D. Dieks (Eds.), *EPSA11 perspectives and foundational problems in philosophy of science* (pp. 115–128). Springer International Publishing.

Bolker, B., Robinson, D., Menne, D., Gabry, J., Buerkner, P., Hau, C., Petry, W., Wiley, J., Kennedy, P., Szöcs, E., Patil, I., Arel-Bundock, V., Denney, B., & Brunson, C. (2024). *broom.mixed: Tidying methods for mixed models* (version 0.2.9.5) [R package]. https://github.com/bbolker/broom.mixed

Borenstein, M., Higgins, J. P. T., Hedges, L., & Rothstein, H. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*, 5–18. https://doi.org/10.1002/jrsm.1230

Borer, E. T., Harpole, W. S., Adler, P. B., Lind, E. M., Orrock, J. L., Seabloom, E. W., & Smith, M. D. (2014). Finding generality in ecology: A model for globally distributed experiments. *Methods in Ecology and Evolution*, *5*(1), 65–73. https://doi.org/10.1111/2041-210X.12125

Borer, E. T., & Seabloom, E. W. (2025). Maximizing inference from distributed experimental networks via "add-on" studies. *Frontiers in Ecology and the Environment*, *23*(10), e70007. https://doi.org/10.1002/fee.70007

Borger, M. J., & Ramesh, A. (2025). Let's DAG in: How directed acyclic graphs can help behavioural ecology be more transparent. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2051), 20250963. https://doi.org/10.1098/rspb.2025.0963

Borregaard, M. K., & Hart, E. M. (2016). Towards a more reproducible ecology. *Ecography*, *39*(4), 349–353. https://doi.org/10.1111/ecog.02493

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M.,

Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Braga, P. H. P., Hébert, K., Hudgins, E. J., Scott, E. R., Edwards, B. P. M., Sánchez Reyes, L. L., Grainger, M. J., Foroughirad, V., Hillemann, F., Binley, A. D., Brookson, C. B., Gaynor, K. M., Shafiei Sabet, S., Güncan, A., Weierbach, H., Gomes, D., & Crystal-Ornelas, R. (2023). Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution. *Methods in Ecology and Evolution*, *14*(6), 1364–1380. https://doi.org/10.1111/2041-210X.14108

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., & Bahnsen, O. (2021). *Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty*.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Briga, M., & Verhulst, S. (2021). Mosaic metabolic ageing: Basal and standard metabolic rates age in opposite directions and independent of environmental quality, sex and life span in a passerine. *Functional Ecology*, *35*(5), 1055–1068. https://doi.org/10.1111/1365-2435.13785

Briscoe, N. J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J. J., Camac, J. S., Giljohann, K. M., Holden, M. H., Hradsky, B. A., Kearney, M. R., McMahon, S. M., Phillips, B. L., Regan, T. J., Rhodes, J. R., Vesk, P. A., Wintle, B. A., Yen, J. D. L., & Guillera-Arroita, G. (2019). Forecasting species range dynamics with process-explicit models: Matching methods to applications. *Ecology Letters*, *22*(11), 1940–1956. https://doi.org/10.1111/ele.13348

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). Do preregistration and preanalysis plans reduce *p* -hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. *Journal of Political Economy Microeconomics*, *2*(3), 527–561. https://doi.org/10.1086/730455

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Brousil, M. R., Filazzola, A., Meyer, M. F., Sharma, S., & Hampton, S. E. (2023). Improving ecological data science with workflow management software. *Methods in Ecology and Evolution*, *14*(6), 1381–1388. https://doi.org/10.1111/2041-210X.14113

Brudvig, L. A. (2017). Toward prediction in the restoration of biodiversity. *Journal of Applied Ecology*, *54*(4), 1013–1017. https://doi.org/10.1111/1365-2664.12940

Burgman, M., Chiaravalloti, R., Fidler, F., Huan, Y., McBride, M., Marcoci, A., Norman, J., Vercammen, A., Wintle, B., & Yu, Y. (2023). A toolkit for open and pluralistic conservation science. *Conservation Letters*, *16*(1), e12919. https://doi.org/10.1111/conl.12919

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretical approach* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/b97636

Bush, M., Holcombe, A. O., Wintle, B. C., Fidler, F., & Vazire, S. (2019). Real problem, wrong solution: Why the nationals shouldn't politicise the science replication crisis. *The Conversation*. https://theconversation.com/real-problem-wrong-solution-why-the-nationals-shouldnt-politicise-the-science-replication-crisis-124076

Butler, N., Delaney, H., & Spoelstra, S. (2017). The gray zone: Questionable research practices in the business school. *Academy of Management Learning & Education*, *16*(1), 94–109. http://www.jstor.org/stable/26400167

Buxton, R. T., Nyboer, E. A., Pigeon, K. E., Raby, G. D., Rytwinski, T., Gallagher, A. J., Schuster, R., Lin, H.-Y., Fahrig, L., Bennett, J. R., Cooke, S. J., & Roche, D. G. (2021). Avoiding wasted research resources in conservation science. *Conservation Science and Practice*, *3*(2), e329. https://doi.org/10.1111/csp2.329

Bylemans, J., Everts, T., Brys, R., & Duncan, R. P. (2025). From anarchy to clarity, data preprocessing and statistical choices influence quantitative environmental DNA analyses. *Methods in Ecology and Evolution*, *16*(7), 1322–1333. https://doi.org/10.1111/2041-210X.70064

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, *96*(9), 2370–2382. https://doi.org/10.1890/14-1639.1

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210x.13559

Capilla-Lasheras, P., Thompson, M. J., Sánchez-Tójar, A., Haddou, Y., Branston, C. J., Réale, D., Charmantier, A., & Dominoni, D. M. (2022). A global meta-analysis reveals higher variation in breeding phenology in urban birds than in their non-urban neighbours. *Ecology Letters*, *25*(11), 2552–2570. https://doi.org/10.1111/ele.14099

Carmel, Y., Kent, R., Bar-Massada, A., Blank, L., Liberzon, J., Nezer, O., Sapir, G., & Federman, R. (2013). Trends in ecological research during the last three decades–a systematic review. *PLOS ONE*, *8*(4), e59813. https://doi.org/10.1371/journal.pone.0059813

Center for Open Science. (2025). *COS statement on "restoring gold standard science" executive order*. [Press release]. https://www.cos.io/about/news/cos-statement-on-restoring-gold-standard-science-executive-order Accessed: 2025-11-26

Christie, A. P., Amano, T., Martin, P. A., Petrovan, S. O., Shackelford, G. E., Simmons, B. I., Smith, R. K., Williams, D. R., Wordley, C. F. R., & Sutherland, W. J. (2021). The challenge of biased evidence in conservation. *Conservation Biology*, *35*(1), 249–262. https://doi.org/10.1111/cobi.13577

Christin, S., Hervet, É., & Lecomte, N. (2020). Going further with model verification and deep learning. *Methods in Ecology and Evolution*, *12*(1), 130–134. https://doi.org/10.1111/2041-210x.13494

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79. https://doi.org/10.1145/3360311

Cologna, V., Mede, N. G., Berger, S., Besley, J., Brick, C., Joubert, M., Maibach, E. W., Mihelj, S., Oreskes, N., Schäfer, M. S., Van Der Linden, S., Abdul Aziz, N. I., Abdulsalam, S., Shamsi, N. A., Aczel, B., Adinugroho, I., Alabrese, E., Aldoh, A., Alfano, M., … Zwaan, R. A. (2025). Trust in scientists and their role in society across 68 countries. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-02090-5

Commander, C. J. C., Barnett, L. A. K., Ward, E. J., Anderson, S. C., & Essington, T. E. (2022). The shadow model: How and why small choices in spatially explicit species distribution models affect predictions. *PeerJ*, *10*, e12783. https://doi.org/10.7717/peerj.12783

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *88*(4), 526–542. https://doi.org/10.1002/ecm.1314

Connolly, S. R., Keith, S. A., Colwell, R. K., & Rahbek, C. (2017). Process, mechanism, and modeling in macroecology. *Trends in Ecology & Evolution*, *32*(11), 835–844. https://doi.org/10.1016/j.tree.2017.08.011

Conroy, M. J., & Peterson, J. T. (2013). *Decision Making in Natural Resource Management: A Structured, Adaptive Approach*. Wiley Blackwell.

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P., Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E. J., Benavides, G. M., Benker, N., BensonMeyer, E. P., … Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162567

Corkery, I., Barea, L., Giejsztowt, J., Maseyk, F., & Mealey, C. (2023). Poorly designed biodiversity loss-gain models facilitate biodiversity loss in new zealand. *New Zealand Journal of Ecology*, *47*(1), 3548. https://doi.org/10.20417/nzjecol.47.3548

Corneille, O., Havemann, J., Henderson, E. L., IJzerman, H., Hussey, I., Orban De Xivry, J.-J., Jussim, L., Holmes, N. P., Pilacinski, A., Beffara, B., Carroll, H., Outa, N. O., Lush, P., & Lotter, L. D. (2023). Beware 'persuasive communication devices' when writing and reading scientific articles. *eLife*, *12*, e88654. https://doi.org/10.7554/eLife.88654

Cuevas Shaw, L., Errington, T. M., & Mellor, D. T. (2022). Toward open science: Contributing to research culture change. *Science Editor*, *45*, 14–17. https://doi.org/10.36591/SE-D-4501-14

Čulina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S., & Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. *Nature Ecology & Evolution*, *2*, 420–426. https://doi.org/10.1038/s41559-017-0458-2

Čulina, A., Berg, I. van den, Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLOS Biology*, *18*(7), e3000763. https://doi.org/10.1371/journal.pbio.3000763

Dancho, M., & Vaughan, D. (2023). *Timetk: A tool kit for working with time series* (version 2.8.3) [R package]. https://CRAN.R-project.org/package=timetk

Davis, A. J., & Kay, S. (2023). Writing statistical methods for ecologists. *Ecosphere*, *14*(5), e4539. https://doi.org/10.1002/ecs2.4539

de Vos, M. G., Janssen, S. J. C., & Van Brussel, L. G. J. (2011). Are environmental models transparent and reproducible enough? *19th International Congress on Modelling and Simulation International Congress on Modelling and Simulation*, 2954–2961. https://www.mssanz.org.au/modsim2011/G7/devos.pdf

DeAngelis, D. L., Franco, D., Hastings, A., Hilker, F. M., Lenhart, S., Lutscher, F., Petrovskaya, N., Petrovskii, S., & Tyson, R. C. (2021). Towards building a sustainable future: Positioning ecological modelling for impact in ecosystems management. *Bulletin of Mathematical Biology*, *83*(10), 107. https://doi.org/10.1007/s11538-021-00927-y

DeKogel, C. H. (1997). Long-term effects of brood size manipulation on morphological development and sex-specific mortality of offspring. *Journal of Animal Ecology*, *66*(2), 167–178. https://doi.org/10.2307/6019

Deressa, T. K., Stern, D. I., Vangronsveld, J., Minx, J., Lizin, S., Malina, R., & Bruns, S. B. (2023). More than half of statistically significant research findings in the environmental sciences are actually not [Preprint]. *EcoEvoRxiv*. https://doi.org/10.32942/X24G6Z

Desbureaux, S. (2021). Subjective modeling choices and the robustness of impact evaluations in conservation science. *Conservation Biology*, *35*(5), 1615–1626. https://doi.org/10.1111/cobi.13728

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805. https://doi.org/10.1098/rsos.200805

Dicks, L. V., Walsh, J. C., & Sutherland, W. J. (2014). Organising evidence for environmental management decisions: a '4S' hierarchy. *Trends in Ecology & Evolution*, *29*(11), 607–613. https://doi.org/10.1016/j.tree.2014.09.004

Dietze, M. C. (2017). Prediction in ecology: A first-principles framework. *Ecological Applications*, *27*(7), 2048–2060.

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018b). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loescher, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., & White, E. P. (2018a). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., & Gerstner, K. (2018). Model averaging in ecology: A review of bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. https://doi.org/10.1002/ecm.1309

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*(781), 108–111. https://doi.org/10.3389/fpsyg.2015.00621

Eker, S., Rovenskaya, E., Obersteiner, M., & Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature Communications*, *9*(1), 5359. https://doi.org/10.1038/s41467-018-07811-9

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Ellison, Aaron M., Kress, W. John., & Shavit, A. (2018). *Stepping in the same river twice : Replication in biological research*. Yale University Press,.

Engelschalt, P., Röske, M., Penzlin, J., Krüger, D., & Upmeier zu Belzen, A. (2023). Abductive reasoning in modeling biological phenomena as complex systems. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1170967

Erickson, R. A., Archer, A. A., & Fienen, M. N. (2024). A reproducible manuscript workflow with a quarto template. *Journal of Fish and Wildlife Management*, *15*(1), 251–258. https://doi.org/10.3996/JFWM-24-003

Essawy, B. T., Goodall, J. L., Voce, D., Morsy, M. M., Sadler, J. M., Choi, Y. D., Tarboton, D. G., & Malik, T. (2020). A taxonomy for reproducible and replicable research in environmental modelling. *Environmental Modelling & Software*, *134*, 104753. https://doi.org/10.1016/j.envsoft.2020.104753

Evans, T. R., Branney, P., Clements, A., & Hatton, E. (2023). Improving evidence-based practice through preregistration of applied research: Barriers and recommendations. *Accountability in Research*, *30*(2), 88–108. https://doi.org/10.1080/08989621.2021.1969233

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719. https://doi.org/10.1073/pnas.1618569114

Fanelli, D., & Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*(37), 15031–15036. https://doi.org/10.1073/pnas.1302997110

Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution*, *3*(10), 1382–1395. https://doi.org/10.1038/s41559-019-0972-5

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology*, *20*(5), 1539–1544. https://doi.org/10.1111/j.1523-1739.2006.00525.x

Fidler, F., Chee, Y. E., Wintle, B. A., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for Evaluating Reproducibility in Ecology and Evolution. *BioScience*, biw159–8. https://doi.org/10.1093/biosci/biw159

Fidler, F., Fraser, H., McCarthy, M. A., & Game, E. T. (2018). Improving the transparency of statistical reporting in Conservation Letters. *Conservation Letters*, *11*(2), e12453–4. https://doi.org/10.1111/conl.12453

Filazzola, A., & Lortie, C. J. (2022). A call for clean code to effectively communicate science. *Methods in Ecology and Evolution*, *13*(10), 2119–2128. https://doi.org/10.1111/2041-210x.13961

Firke, S. (2023). *janitor: Simple tools for examining and cleaning dirty data* (version 2.2.0) [R package]. https://github.com/sfirke/janitor

Fitzpatrick, M. C., Lachmuth, S., & Haydt, N. T. (2021). The ODMAP protocol: A new tool for standardized reporting that could revolutionize species distribution modeling. *Ecography*, *44*(7), 1067–1070. https://doi.org/10.1111/ecog.05700

Flournoy, J. C., Vijayakumar, N., Cheng, T. W., Cosme, D., Flannery, J. E., & Pfeifer, J. H. (2020). Improving practices and inferences in developmental cognitive neuroscience. *Developmental Cognitive Neuroscience*, *45*, 100807. https://doi.org/10.1016/j.dcn.2020.100807

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, *65*(1), 47–55. https://doi.org/10.1007/s00265-010-1038-5

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. https://doi.org/10.1111/geb.12684

Franks, D. W., Ruxton, G. D., & Sherratt, T. (2025). Ecology needs a causal overhaul. *Biological Reviews*, *100*(5), 1950–1969. https://doi.org/10.1111/brv.70029

Fraser, H., Barnett, A., Parker, T. H., & Fidler, F. (2020). The role of replication studies in ecology. *Ecology and Evolution*, *10*(12), 5197–5207. https://doi.org/10.1002/ece3.6330

Fraser, H., Parker, T. H., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Freckleton, R. P. (2018). Accessibility, reusability, reliability: Improving the standards for publishing code in *Methods in Ecology and Evolution*. *Methods in Ecology and Evolution*, *9*(1), 4–6. https://doi.org/10.1111/2041-210X.12947

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement* (version 0.84.1) [R package]. https://CRAN.R-project.org/package=irr

Garbin, C., & Marques, O. (2022). Assessing methods and tools to improve reporting, increase transparency, and reduce failures in machine learning applications in health care. *Radiology: Artificial Intelligence*, *4*(2), e210127. https://doi.org/10.1148/ryai.210127

García-Díaz, P., Prowse, T. A. A., Anderson, D. P., Lurgi, M., Binny, R. N., & Cassey, P. (2019). A concise guide to developing and using quantitative models in conservation management. *Conservation Science and Practice*, *1*(2), e11. https://doi.org/10.1002/csp2.11

Gardner, C. J., Waeber, P. O., Razafindratsima, O. H., & Wilmé, L. (2018). Decision complacency and conservation planning. *Conservation Biology*, *32*(6), 1469–1472. https://doi.org/10.1111/cobi.13124

Geary, W. L., Bode, M., Doherty, T. S., Fulton, E. A., Nimmo, D. G., Tulloch, A. I. T., Tulloch, V. J. D., & Ritchie, E. G. (2020). A guide to ecosystem models and their

environmental applications. *Nature Ecology & Evolution*, *4*(11), 1459–1471. https://doi.org/10.1038/s41559-020-01298-8

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*(4), 310–316. http://www.jstor.org/stable/27859361

Getz, W. M., Marshall, C. R., Carlson, C. J., Giuggioli, L., Ryan, S. J., Romañach, S. S., Boettiger, C., Chamberlain, S. D., Larsen, L., D'Odorico, P., & O'Sullivan, D. (2017). Making ecological models adequate. *Ecology Letters*, *21*(2), 153–166. https://doi.org/10.1111/ele.12893

Giljohann, K. M., McCarthy, M. A., & Kelly, L. T. (2015). Choice of biodiversity index drives optimal fire management decisions. *Ecological Applications*, *25*(1), 264–277. https://doi.org/10.1890/14-0257.1

Gomes, D. (2025). How will we prepare for an uncertain future? The value of open data and code for unborn generations facing climate change. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2040), 20241515. https://doi.org/10.1098/rspb.2024.1515

Gomes, D., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., Sánchez-Reyes, L. L., Turba, R., Martinez, P. A., Moreau, D., Bertram, M. G., Smout, C. A., & Gaynor, K. M. (2022). *Why don't we share data and code? Perceived barriers and benefits to public archiving practices*. *289*(1987), 20221113. https://doi.org/10.1098/rspb.2022.1113

Gould, E. (2026). *EcoConsModPreReg: A quarto template for preregistering ecological modelling studies* (v.0.0.9005). Zenodo. https://doi.org/10.5281/zenodo.18907767

Gould, E., Fraser, H. S., Nakagawa, S., & Parker, T. H. (2023). *ManyEcoEvo: An r package for analysing many-analyst style data.* (v2.7.6) [R package]. Zenodo. https://doi.org/10.5281/zenodo.10046153

Gould, E., Fraser, H. S., Parker, T. H., & Nakagawa, S. (2024). Egouldo/ManyAnalysts: Manuscript source code for "same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology" [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.13850927

Gould, E., Fraser, H., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker, C. M., Barrett, M., … Zitomer, R. A. (2025). Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, *23*(1), 35. https://doi.org/10.1186/s12915-024-02101-x

Gould, E., Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H. F., Good, M. K., Duncan, D. H., Hauser, C. E., Wintle, B. C., & Rumpff, L. (2026). "But I can't preregister my research": Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research. *Methods in Ecology and Evolution*, 2041–210x.70311. https://doi.org/10.1111/2041-210x.70311

Gould, E., Jones, C., Yen, J. D. L., Fraser, H., Wootton, H., Vivian, L., Good, M., Duncan, D., Rumpff, L., & Fidler, F. (2026). *EcoConsPreReg: A guide to adaptive preregistration for model-based research in ecology and conservation*. https://doi.org/10.5281/zenodo.19064144

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129–139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., & Berger, U. (2016). Robustness analysis: Deconstructing computational models for ecological theory and applications. *Ecological Modelling*, *326*, 162–167. https://doi.org/10.1016/j.ecolmodel.2015.07.018

Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., Johnston, A. S. A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwäder, M.-S., Stillman, R. A., Thiele, J. C., & Ayllón, D. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal*

of Artificial Societies and Social Simulation, *23*(2), 7. https://doi.org/10.18564/jasss.4259

Grueber, C. E., Nakagawa, S., Laws, R. J., & Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: Challenges and solutions. *Journal of Evolutionary Biology*, *24*(4), 699–711. https://doi.org/doi:10.1111/j.1420-9101.2010.02210.x

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Haddaway, N. R., Macura, B., Whaley, P., & Pullin, A. S. (2018). ROSES RepOrting standards for systematic evidence syntheses: Pro forma, flow-diagram and descriptive summary of the plan and conduct of environmental systematic reviews and systematic maps. *Environmental Evidence*, *7*(7). https://doi.org/10.1186/s13750-018-0121-7

Hämäläinen, R. P., & Lahtinen, T. J. (2016). Path dependence in operational research—how the modeling process can influence the results. *Operations Research Perspectives*, *3*, 14–20. https://doi.org/10.1016/j.orp.2016.03.001

Hamilton, S. H., Fu, B., Guillaume, J. H. A., Badham, J., Elsawah, S., Gober, P., Hunt, R. J., Iwanaga, T., Jakeman, A. J., Ames, D. P., Curtis, A., Hill, M. C., Pierce, S. A., & Zare, F. (2019). A framework for characterising and evaluating the effectiveness of environmental modelling. *Environmental Modelling & Software*, *118*, 83–98. https://doi.org/10.1016/j.envsoft.2019.04.008

Hamilton, S. H., Pollino, C. A., Stratford, D. S., Fu, B., & Jakeman, A. J. (2022). Fit-for-purpose environmental modeling: Targeting the intersection of usability, reliability and feasibility. *Environmental Modelling & Software*, *148*, 105278. https://doi.org/10.1016/j.envsoft.2021.105278

Haroz, S. (2022, February 24). *Comparison of Preregistration Platforms*. https://doi.org/10.31222/osf.io/zry2u

Harrell Jr, F. E. (2024). *Hmisc: Harrell miscellaneous* (version 5.1-2) [R package]. https://hbiostat.org/R/Hmisc/

Hemming, V., Camaclang, A. E., Adams, M. S., Burgman, M., Carbeck, K., Carwardine, J., Chadès, I., Chalifour, L., Converse, S. J., Davidson, L. N. K., Garrard, G. E., Finn, R., Fleri, J. R., Huard, J., Mayfield, H. J., Madden, E. M., Naujokaitis-Lewis, I., Possingham, H. P., Rumpff, L., … Martin, T. G. (2021). An introduction to decision science for conservation. *Conservation Biology*, *36*(1), e13868. https://doi.org/10.1111/cobi.13868

Herman, B. D., McKay, S. K., Altman, S., Richards, N. S., Reif, M., Piercy, C. D., & Swannack, T. M. (2019). Unpacking the black box: Demystifying ecological models through interactive workshops and hands-on learning. *Frontiers in Environmental Science*, *7*. https://doi.org/10.3389/fenvs.2019.00122

Hester, J., Henry, L., Müller, K., Ushey, K., Wickham, H., & Chang, W. (2024). *withr: Run code "With" temporarily modified global state* (version 3.0.0) [R package]. https://withr.r-lib.org

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hildebrandt, M. (2018). Preregistration of machine learning research design against p-hacking. In Emre Bayamlıoğlu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *BEING PROFILED:COGITAS ERGO SUM*. Amsterdam University Press. https://ssrn.com/abstract=3256146

Hillebrand, H., & Gurevitch, J. (2013). Reporting standards in experimental studies. *Ecology Letters*, *16*(12), 1419–1420. https://doi.org/10.1111/ele.12190

Hindman, M. (2015). Building better models. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62. https://doi.org/10.1177/0002716215570279

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Hofman, J. M., Chatzimparmpas, A., Sharma, A., Watts, D. J., & Hullman, J. (2023). Pre-registration for predictive modeling [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2311.18807

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O. K., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., … Wagenmakers, E.-J. (2023). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, *13*(3), 237–283. https://doi.org/10.1080/2153599X.2022.2070255

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, *119*, 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036

Houlahan, J. E., McKinney, S. T., Anderson, T. M., & McGill, B. J. (2016). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1–7. https://doi.org/10.1111/oik.03726

Hulme, P. E. (2014). Bridging the knowing-doing gap: Know-who, know-what, know-why, know-how and know-when. *Journal of Applied Ecology*, *51*(5), 1131–1136. https://doi.org/10.1111/1365-2664.12321

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Hurley, A. G., Peters, R. L., Pappas, C., Steger, D. N., & Heinrich, I. (2022). Addressing the need for interactive, efficient, and reproducible data processing in ecology with the datacleanr r package. *PLOS ONE*, *17*(5), e0268426. https://doi.org/10.1371/journal.pone.0268426

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., & Seo, J. (2024). *gt: Easily create presentation-ready display tables* (version 0.10.1) [R package]. https://gt.rstudio.com

Iannone, R., Vargas, M., & Choe, J. (2025). *Pointblank: Data validation and organization of metadata for local and remote tables*. https://doi.org/10.32614/CRAN.package.pointblank R package version 0.12.3

Ioannidis, J. P. A. (2022). Pre-registration of mathematical models. *Mathematical Biosciences*, *345*, 108782. https://doi.org/10.1016/j.mbs.2022.108782

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Ivimey-Cook, E. R., Sánchez-Tójar, A., Berberi, I., Čulina, A., Roche, D. G., Almeida, R. A., Amin, B., Bairos-Novak, K. R., Balti, H., Bertram, M., Bliard, L., Byrne, I., Chan, Y.-C., Cioffi, W. R., Corbel, Q., Elsy, A. D., Florko, K. R. N., Gould, E., Grainger, M., …

Moran, N. P. (2025). From policy to practice: Progress towards data- and code-sharing in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences*, *292*(2055), 20251394. https://doi.org/10.1098/rspb.2025.1394

Jakeman, A. J., Elsawah, S., Wang, H.-H., Hamilton, S. H., Melsen, L., & Grimm, V. (2024). Towards normalizing good practice across the whole modeling cycle: Its instrumentation and future research topics. *Socio-Environmental Systems Modelling*, *6*, 18755. https://doi.org/10.18174/sesmo.18755

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Jakeman, A. J., Voinov, A. A., Rizzoli, A. E., & Chen, S. H. (2008). *Environmental modelling, software and decision support: State of the art and new perspective*. Elsevier.

Janssen, M. A., Pritchard, C., & Lee, A. (2020). On code sharing and model documentation of published individual and agent-based models. *Environmental Modelling & Software*, *134*, 104873. https://doi.org/10.1016/j.envsoft.2020.104873

Jennions, M. D., Lortie, C. J., Rosenberg, M. S., & Rothstein, H. R. (2013). Publication and related biases. In J. Koricheva, J. Gurevitch, & K. Mengersen (Eds.), *Handbook of meta-analysis in ecology and evolution* (pp. 207–236). Princeton University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Jones, C. S., Lyndsey, V. L., Wootton, H., Gould, E., Yen, J. D. L., Mole, B., Thomas, F., Backstrom, A., & Just, K. (2025). *Vegetation responses to managed river flow events and regimes* [Unpublished Report]. Arthur Rylah Institute for Environmental Research; Department of Energy, Environment and Climate Action.

Jones, C. S., Vivian, V. L., Wootton, H., Gould, E., Yen, J. D. L., Mole, B., Thomas, F., Sutton, F., Crowther, D., Backstrom, A., & Just, K. (2026, April 2). *Vegetation responses to managed river flow events and regimes*. https://doi.org/10.32942/X2S67G

Jones, C. S., Yen, J. D. L., Gould, E., Wootton, H., & Vivian, L. (2026). *Analysis Code and Data for: "Vegetation responses to managed river flow events and regimes"* (0.21.2). Zenodo. https://doi.org/10.5281/ZENODO.19064000

Josefsson, J., Hiron, M., Arlt, D., Auffret, A. G., Berg, Å., Chevalier, M., Glimskär, A., Hartman, G., Kačergytė, I., Klein, J., Knape, J., Laugen, A. T., Low, M., Paquet, M.,

Pasanen-Mortensen, M., Rosin, Z. M., Rubene, D., Żmihorski, M., & Pärt, T. (2020). Improving scientific rigour in conservation evaluations and a plea deal for transparency on potential biases. *Conservation Letters*, *13*(5), e12726. https://doi.org/10.1111/conl.12726

Kambouris, S., Wilkinson, D. P., Smith, E. T., & Fidler, F. (2024). Computationally reproducing results from meta-analyses in ecology and evolutionary biology using shared code and data. *PLOS ONE*, *19*(3), e0300333. https://doi.org/10.1371/journal.pone.0300333

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kass, J. M., Muscarella, R., Galante, P. J., Bohl, C. L., Pinilla-Buitrago, G. E., Boria, R. A., Soley-Guardia, M., & Anderson, R. P. (2021). ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. *Methods in Ecology and Evolution*, *12*(9), 1602–1608. https://doi.org/10.1111/2041-210X.13628

Kass, J. M., Smith, A. B., Warren, D. L., Vignali, S., Schmitt, S., Aiello-Lammens, M. E., Arlé, E., Márcia Barbosa, A., Broennimann, O., Cobos, M. E., Guéguen, M., Guisan, A., Merow, C., Naimi, B., Nobis, M. P., Ondo, I., Osorio-Olvera, L., Owens, H. L., Pinilla-Buitrago, G. E., … Zurell, D. (2025). Achieving higher standards in species distribution modeling by leveraging the diversity of available software. *Ecography*, *2025*(2), e07346. https://doi.org/10.1111/ecog.07346

Kassambara, A. (2023). *ggpubr: "ggplot2" based publication ready plots* (version 0.6.0) [R package]. https://rpkgs.datanovia.com/ggpubr

Kearney, M., McGeoch, M., & Chown, S. (2019). Where do functional traits come from? The role of theory and models. *Biodiversity Information Science and Standards*, *3*, e37500. https://doi.org/10.3897/biss.3.37500

Kellner, K. F., Doser, J. W., & Belant, J. L. (2024). Functional r code is rare in species distribution and abundance papers. *Ecology*, *106*(1), e4475. https://doi.org/10.1002/ecy.4475

Kim, A. Y., Herrmann, V., Barreto, R., Calkins, B., Gonzalez-Akre, E., Johnson, D. J., Jordan, J. A., Magee, L., McGregor, I. R., Montero, N., Novak, K., Rogers, T., Shue, J., & Anderson-Teixeira, K. J. (2022). Implementing GitHub actions continuous integration to reduce

error rates in ecological data collection. *Methods in Ecology and Evolution*, *13*(11), 2572–2585. https://doi.org/10.1111/2041-210X.13982

Kimmel, K., Avolio, M. L., & Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, *7*(9), 1525–1536. https://doi.org/10.1038/s41559-023-02144-3

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Knief, U., & Forstmeier, W. (2025). Are our 95% CIs only worth 45% confidence? [Preprint]. *bioRxiv*. https://doi.org/10.1101/2025.08.05.668631

Knight, A. T., Cowling, R. M., Rouget, M., Balmford, A., Lombard, A. T., & Campbell, B. M. (2008). Knowing but not doing: Selecting priority conservation areas and the research-implementation gap. *Conservation Biology*, *22*(3), 610–617. https://doi.org/10.1111/j.1523-1739.2008.00914.x

Knight, K. (2000). *Mathematical statistics*. Chapman and Hall.

Koivisto, E., & Mäntylä, E. (2024). Are open science instructions targeted to ecologists and evolutionary biologists sufficient? A literature review of guidelines and journal data policies. *Ecology and Evolution*, *14*(7), e11698. https://doi.org/10.1002/ece3.11698

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*(4), 828–844. https://doi.org/10.1111/1365-2745.12224

Kossmeier, M., Tran, U. S., & Voracek, M. (2020). *Metaviz: Forest plots, funnel plots, and visual funnel plot inference for meta-analysis*. https://doi.org/10.32614/CRAN.package.metaviz R package version 0.3.1

Kou-Giesbrecht, S., & Menge, D. N. L. (2021). Nitrogen-fixing trees increase soil nitrous oxide emissions: A meta-analysis. *Ecology*, *102*(8), e03415. https://doi.org/10.1002/ecy.3415

Koziol, M. (2025). 'Your countries are going to hell': Trump stuns UN with tirade on immigration, climate 'con job'. *The Age*. https://www.theage.com.au/world/north-america/trump-stuns-un-with-tirade-on-immigration-climate-hoax-20250924-p5mxg0.html

Kuhn, M., & Frick, H. (2022). *multilevelmod: Model wrappers for multi-level models* (version 1.0.0) [R package]. https://github.com/tidymodels/multilevelmod

Kuhn, M., & Vaughan, D. (2024). *Parsnip: A common API to modeling and analysis functions* (version 1.2.1) [R package]. https://CRAN.R-project.org/package=parsnip

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* [R packages]. https://www.tidymodels.org

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lahtinen, T. J., Guillaume, J. H. A., & Hämäläinen, R. P. (2017). Why pay attention to paths in the practice of environmental modelling. *Environmental Modelling & Software*, *92*, 74–81. https://doi.org/10.1016/j.envsoft.2017.02.019

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1), 117094. https://doi.org/10.1525/collabra.117094

Lakens, D., & DeBruine, L. M. (2021). Improving Transparency, Falsifiability, and Rigor by Making Hypothesis Tests Machine-Readable. *Advances in Methods and Practices in Psychological Science*, *4*(2), 2515245920970949. https://doi.org/10.1177/2515245920970949

Landau, W. M. (2021). The targets r package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

Law, E. A., Ferraro, P. J., Arcese, P., Bryan, B. A., Davis, K., Gordon, A., Holden, M. H., Iacona, G., Martinez, R. M., McAlpine, C. A., Rhodes, J. R., Sze, J. S., & Wilson, K. A. (2017). Projecting the performance of conservation interventions. *Biological Conservation*, *215*, 142–151. https://doi.org/10.1016/j.biocon.2017.08.029

Letcher, R. A. K., Jakeman, A. J., Barreteau, O., Borsuk, M. E., ElSawah, S., Hamilton, S. H., Henriksen, H. J., Kuikka, S., Maier, H. R., Rizzoli, A. E., Delden, H. van, & Voinov, A. A. (2013). Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental Modelling & Software*, *47*(C), 159–181. https://doi.org/10.1016/j.envsoft.2013.05.005

Lewis, A. S. L., Rollinson, C. R., Allyn, A. J., Ashander, J., Brodie, S., Brookson, C. B., Collins, E., Dietze, M. C., Gallinat, A. S., Juvigny-Khenafou, N., Koren, G., McGlinn, D. J., Moustahfid, H., Peters, J. A., Record, N. R., Robbins, C. J., Tonkin, J., & Wardle, G. M. (2023). The power of forecasts to advance ecological theory. *Methods in Ecology and Evolution*, *14*(3), 746–756. https://doi.org/10.1111/2041-210X.13955

Leybourne, D. J., Preedy, K. F., Valentine, T. A., Bos, J. I. B., & Karley, A. J. (2021). Drought has negative consequences on aphid fitness and plant vigor: Insights from a meta-analysis. *Ecology and Evolution*, *11*(17), 11915–11929. https://doi.org/10.1002/ece3.7957

Liu, Y., Althoff, T., & Heer, J. (2020). Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. *Proceedings of the ACM Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3313831.3376533

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2021). Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1753–1763. https://doi.org/10.1109/tvcg.2020.3028985

Lortie, C. J., Aarssen, L. W., Budden, A. E., Koricheva, J. K., Leimu, R., & Tregenza, T. (2007). Publication bias and merit in ecology. *Oikos*, *116*(7), 1247–1253. https://doi.org/10.1111/j.0030-1299.2007.15686.x

Lortie, C. J., & Filazzola, A. (2020). A contrast of meta and metafor packages for meta-analyses in r. *Ecology and Evolution*, *10*(20), 10916–10921. https://doi.org/10.1002/ece3.6747

Lu, X., & White, H. (2014). Robustness checks and robustness tests in applied economics. *Journal of Econometrics*, *178*, 194–206. https://doi.org/10.1016/j.jeconom.2013.08.016

Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, *90*(4), e01422. https://doi.org/10.1002/ecm.1422

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Lüdecke, D., Patil, I., Ben-Shachar, M. S., Wiernik, B. M., Waggoner, P., & Makowski, D. (2021). see: An R package for visualizing statistical models. *Journal of Open Source Software*, *6*(64), 3393. https://doi.org/10.21105/joss.03393

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency: introduction. *PS: Political Science & Politics*, *47*(1), 19–42. https://doi.org/10.1017/S1049096513001716

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodiversity and Conservation*, *9*(5), 655671. https://doi.org/10.1023/a:1008985925162

MacEachern, S. N., & Van Zandt, T. (2019). Preregistration of modeling exercises may not be useful. *Computational Brain & Behavior*, *2*(3-4), 179–182. https://doi.org/10.1007/s42113-019-00038-x

Maitner, B., Santos Andrade, P. E., Lei, L., Kass, J., Owens, H. L., Barbosa, G. C. G., Boyle, B., Castorena, M., Enquist, B. J., Feng, X., Park, D. S., Paz, A., Pinilla-Buitrago, G., Merow, C., & Wilson, A. (2024). Code sharing in ecology and evolution increases citation rates but remains uncommon. *Ecology and Evolution*, *14*(8), e70030. https://doi.org/10.1002/ece3.70030

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. A. (2021). Both questionable and open research practices are prevalent in education research. *Educational Researcher*, *50*(8), 493–504. https://doi.org/10.3102/0013189X211001356

Makowski, D., Ben-Shachar, M. S., Wiernik, B. M., Patil, I., Thériault, R., & Lüdecke, D. (2025). modelbased: An R package to make the most out of your statistical models through marginal means, marginal effects, and model predictions. *Journal of Open Source Software*, *10*(109), 7969. https://doi.org/10.21105/joss.07969

ManyBirds Project, Miller, R., Šlipogor, V., Caspar, K. R., Lois-Milevicich, J., Soulsbury, C., Reber, S. A., Mettke-Hofmann, C., Lambert, M., Ashton, B. J., Auersperg, A. M. I., Bateson, M., Belle, S., Bilčík, B., Biondi, L. M., Bonadonna, F., Brucks, D., Butler, M. W., Caro, S. P., … Zanutto, J. A. (2025). A large-scale study across the avian clade identifies ecological drivers of neophobia. *PLOS Biology*, *23*(10), e3003394. https://doi.org/10.1371/journal.pbio.3003394

Marshall, B. M., & Duthie, A. B. (2024). *A habitat selection multiverse reveals largely consistent results despite a multitude of analysis options* [Preprint]. bioRxiv. https://doi.org/10.1101/2024.06.19.599733

Masur, P. K., & Scharkow, M. (2020). *specr: Conducting and visualizing specification curve analyses* (Version 1.0.0) [R package]. https://CRAN.R-project.org/package=specr

McCrea, R., King, R., Graham, L., & Börger, L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, *14*(1), 4–11. https://doi.org/10.1111/2041-210X.14050

McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., & Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, *13*(586), eabb1655. https://doi.org/10.1126/scitranslmed.abb1655

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLOS ONE*, *10*(8), e0136088. https://doi.org/10.1371/journal.pone.0136088

McIntire, E. J. B., Chubaty, A. M., Cumming, S. G., Andison, D., Barros, C., Boisvenue, C., Haché, S., Luo, Y., Micheletti, T., & Stewart, F. E. C. (2022). PERFICT: A re-imagined

foundation for predictive ecology. *Ecology Letters*, *25*(6), 1345–1351. https://doi.org/10.1111/ele.13994

McMillan, B. T. M., Baumgartner, H. A., Bergmann, C., Frank, M. C., Hamlin, J. K., Kampis, D., Kline Struhl, M., Ko, E.-S., Kosie, J. E., Lew-Williams, C., Lucca, K., Schuwerk, T., Soderstrom, M., Visser, I., Yuen, F. L., Zettersten, M., & Byers-Heinlein, K. (2026). What 5000 babies can tell us about developing minds and how to study them. *Communications Psychology*, *4*(1), 92. https://doi.org/10.1038/s44271-026-00477-w

Meding, K., & Hagendorff, T. (2024). Fairness hacking: The malicious practice of shrouding unfairness in algorithms. *Philosophy & Technology*, *37*(1). https://doi.org/10.1007/s13347-023-00679-8

Meschiari, S. (2022). *latex2exp: Use LaTeX expressions in plots* (version 0.9.6) [R package]. https://www.stefanom.io/latex2exp

Micheletti, T., Wimmler, M.-C., Berger, U., Grimm, V., & McIntire, E. J. (2024). Beyond guides, protocols and acronyms: Adoption of good modelling practices depends on challenging academia's status quo in ecology. *Ecological Modelling*, *496*, 110829. https://doi.org/10.1016/j.ecolmodel.2024.110829

Miles, C. (2008). Testing market-based instruments for conservation in northern victoria. In T. Norton, T. Lefroy, K. Bailey, & G. Unwin (Eds.), *Biodiversity: Integrating conservation and production: Case studies from australian farms, forests and fisheries* (pp. 133–146). CSIRO Publishing.

Millard, S. P. (2013). *EnvStats: An r package for environmental statistics* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-8456-1

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Molina, I., & Marhuenda, Y. (2015). sae: An R package for small area estimation. *The R Journal*, *7*(1), 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

Morrison, C., Wardle, C., & Castley, J. G. (2016). Repeatability and Reproducibility of Population Viability Analysis (PVA) and the Implications for Threatened Species Management. *Frontiers in Ecology and Evolution*, *4*, 875–877. https://doi.org/10.3389/fevo.2016.00098

Morrissey, M. B., & Ruxton, G. D. (2018). Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, *10*(3). https://doi.org/10.3998/ptpbio.16039257.0010.003

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G. J., Lavorel, S., Le Gall, L., Meslin, L., … Loreau, M. (2015). REVIEW: Predictive ecology in a changing world. *Journal of Applied Ecology*, *52*(5), 1293–1310. https://doi.org/doi.org/10.1111/1365-2664.12482

Müller, K. (2020). *here: A simpler way to find your files* (version 1.0.1) [R package]. https://here.r-lib.org/

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Naas, A. E., Keetz, L. T., Halvorsen, R., Horvath, P., Mienna, I. M., Simensen, T., & Bryn, A. (2024). Choice of predictors and complexity for ecosystem distribution models: Effects on performance and transferability. *Ecography*, *2024*(8), e07269. https://doi.org/10.1111/ecog.07269

Nagy, T., Hergert, J., Elsherif, M. M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., Wang, Y. A., Scharfenberg, D., Tyson, G., Yang, Y.-F., Skvortsova, A., Alarie, S., Graves, K., Sotola, L. K., Moreau, D., & Rubínová, E. (2025). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science*, *8*(3). https://doi.org/10.1177/25152459251348431

Nakagawa, S., Armitage, D. W., Froese, T., Yang, Y., & Lagisz, M. (2025). Poor hypotheses and research waste in biology: Learning from a theory crisis in psychology. *BMC Biology*, *23*(1), 33. https://doi.org/10.1186/s12915-025-02134-w

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nakagawa, S., Ivimey-Cook, E. R., Grainger, M. J., O'Dea, R. E., Burke, S., Drobniak, S. M., Gould, E., Macartney, E. L., Martinig, A. R., Morrison, K., Paquet, M., Pick, J. L., Pottier,

P., Ricolfi, L., Wilkinson, D. P., Willcox, A., Williams, C., Wilson, L. A. B., Windecker, S. M., … Lagisz, M. (2023). Method reporting with initials for transparency (MeRIT) promotes more granularity and accountability for author contributions. *Nature Communications*, *14*(1788). https://doi.org/10.1038/s41467-023-37039-1

Nakagawa, S., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Sánchez-Tójar, A., Yang, Y., & O'Dea, R. E. (2022). Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution*, *13*(1), 4–21. https://doi.org/10.1111/2041-210X.13724

Nakagawa, S., Lagisz, M., O'Dea, R. E., Pottier, P., Rutkowska, J., Senior, A. M., Yang, Y., & Noble, D. W. A. (2023). orchaRd 2.0: An r package for visualising meta-analyses with orchard plots. *Methods in Ecology and Evolution*, *14*(8), 2003–2010. https://doi.org/https://doi.org/10.1111/2041-210X.14152

Nakagawa, S., Noble, D. W., Senior, A. M., & Lagisz, M. (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology*, *15*(1), 18. https://doi.org/10.1186/s12915-017-0357-7

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*(88). https://doi.org/10.1186/s12915-015-0196-3

Neoh, M. J. Y., Carollo, A., Lee, A., & Esposito, G. (2023). Fifty years of research on questionable research practises in science: Quantitative analysis of co-citation patterns. *Royal Society Open Science*, *10*(10), 230677. https://doi.org/10.1098/rsos.230677

Nicolaus, M., Michler, S. P. M., Ubels, R., van der Velde, M., Komdeur, J., Both, C., & Tinbergen, J. M. (2009). Sex-specific effects of altered competition on nestling growth and survival: An experimental manipulation of brood size and sex ratio. *Journal of Animal Ecology*, *78*(2), 414–426. https://doi.org/10.1111/j.1365-2656.2008.01505.x

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, *5*, e21451. https://doi.org/10.7554/eLife.21451.001

Noble, D. W. A., Lagisz, M., O'Dea, R. E., & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta-analyses. *Molecular Ecology*, *26*(9), 2410–2425. https://doi.org/10.1111/mec.14031

Nosek, B. A. (2025). Science becomes trustworthy by constantly questioning itself. *PLOS Biology*, *23*(8), e3003334. https://doi.org/10.1371/journal.pbio.3003334

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect Psychol Sci*, *7*(6), 615–631.

O'Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., & Nakagawa, S. (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A PRISMA extension. *Biological Reviews of the Cambridge Philosophical Society*, *96*(5), 1695–1722. https://doi.org/10.1111/brv.12721

O'Dea, R. E., Parker, T. H., Chee, Y. E., Čulina, A., Drobniak, S. M., Duncan, D. H., Fidler, F., Gould, E., Ihle, M., Kelly, C. D., Lagisz, M., Roche, D. G., Sánchez-Tójar, A., Wilkinson, D. P., Wintle, B. C., & Nakagawa, S. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*(1), 68. https://doi.org/10.1186/s12915-021-01006-3

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122. https://doi.org/10.1111/j.2041-210x.2010.00021.x

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Ottaviani, G., Martínez, A., Bon, M. P., & Mammola, S. (2025). *On the quest for novelty in ecology* [Preprint]. https://doi.org/10.1101/2023.02.27.530333

Page, M. J., & Moher, D. (2017). Evaluations of the uptake and impact of the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement and extensions: A scoping review. *Systematic Reviews*, *6*(1), 263. https://doi.org/10.1186/s13643-017-0663-8

Paniw, M., Cousens, R. D., Baker, C., & Le, T. (2023). Theory, prediction and application. In R. D. Cousens (Ed.), *Effective ecology* (1st ed., p. 23). CRC Press. https://doi.org/10.1201/9781003314332

Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., Kelly, C. D., Gurevitch, J., & Nakagawa, S. (2016). Transparency in Ecology and Evolution: Real Problems, Real Solutions. *Trends in Ecology & Evolution*, *31*(9), 711–719. https://doi.org/10.1016/j.tree.2016.07.002

Parker, T. H., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, *33*(4), 747–750. https://doi.org/10.1111/cobi.13342

Parker, T. H., Main, E., Nakagawa, S., Gurevitch, J., Jarrad, F., & Burgman, M. (2016). Promoting transparency in conservation science. *Conservation Biology*, *30*(6), 1149–1150. https://doi.org/10.1111/cobi.12760

Parker, T. H., & Yang, Y. (2023). Exaggerated effects in ecology. *Nature Ecology & Evolution*, *7*(9), 1356–1357. https://doi.org/10.1038/s41559-023-02156-z

Pavelin, K., Cham, J. A., de Matos, P., Brooksbank, C., Cameron, G., & Steinbeck, C. (2012). *PLOS Computational Biology*, *8*(7), 1–4. https://doi.org/10.1371/journal.pcbi.1002554

Pedersen, T. L. (2024). *patchwork: The composer of plots* (version 1.2.0) [R package]. https://patchwork.data-imaginist.com

Pei, Y., Forstmeier, W., Wang, D., Martin, K., Rutkowska, J., & Kempenaers, B. (2020). Proximate causes of infertility and embryo mortality in captive zebra finches. *The American Naturalist*, *196*(5), 577–596. https://doi.org/10.1086/710956

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Pick, J., Allen, B., Bachelot, B., Bairos-Novak, K., Brand, J., Class, B., Dallas, T., D'Amelio, P., Fenollosa, E., Fernández-Juricic, E., Gomes, D., Grainger, M., Guillemaud, T., John, C., Krasnow, R., Lagisz, M., Lequime, S., Maynard, D., Nakagawa, S., … Ivimey-Cook, E. R. (2025). The SORTEE guidelines for data and code quality control in ecology and evolutionary biology [Preprint]. *EcoEvoRxiv*. https://doi.org/10.32942/X24P8S

Planque, B., Aarflot, J. M., Buttay, L., Carroll, J., Fransner, F., Hansen, C., Husson, B., Langangen, Ø., Lindstrøm, U., Pedersen, T., Primicerio, R., Sivel, E., Skogen, M. D., Strombom, E., Stige, L. C., Varpe, Ø., & Yoccoz, N. G. (2022). A standard protocol for describing the evaluation of ecological models. *Ecological Modelling*, *471*, 110059. https://doi.org/10.1016/j.ecolmodel.2022.110059

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019). Ecological data should not be so hard to find and reuse. *Trends in Ecology & Evolution*, *34*(6), 494–496. https://doi.org/10.1016/j.tree.2019.04.005

Poliseli, L., Coutinho, J. G. E., Viana, B., Russo, F., & El-Hani, C. N. (2022). Philosophy of science in practice in ecological model building. *Biology & Philosophy*, *37*(4), 21. https://doi.org/10.1007/s10539-022-09851-4

Popovic, G., Mason, T. J., Drobniak, S. M., Marques, T. A., Potts, J., Joo, R., Altwegg, R., Burns, C. C. I., McCarthy, M. A., Johnston, A., Nakagawa, S., McMillan, L., Devarajan, K., Taggart, P. L., Wunderlich, A., Mair, M. M., Martínez-Lanfranco, J. A., Lagisz, M., & Pottier, P. (2024). Four principles for improved statistical ecology. *Methods in Ecology and Evolution*, *15*(2), 266–281. https://doi.org/10.1111/2041-210X.14270

Powers, S. M., & Hampton, S. E. (2018). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*(1), e01822. https://doi.org/10.1002/eap.1822

Prosperi, M., Bian, J., Buchan, I. E., Koopman, J. S., Sperrin, M., & Wang, M. (2019). Raiders of the lost HARK: A reproducible inference framework for big data science. *Palgrave Communications*, *5*(125). https://doi.org/10.1057/s41599-019-0340-8

Pu, X., Zhu, L., Kay, M., & Conrad, F. (2019). Designing for preregistration in practice: Multiple norms and purposes. *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'19 Extended Abstracts)*, 7. https://doi.org/10.1145/3290607.3312862

Purgar, M., Glasziou, P., Klanjscek, T., Nakagawa, S., & Čulina, A. (2024). Supporting study registration to reduce research waste. *Nature Ecology & Evolution*, *8*(8), 1391–1399. https://doi.org/10.1038/s41559-024-02433-5

Purgar, M., Klanjscek, T., & Čulina, A. (2022). Quantifying research waste in ecology. *Nature Ecology & Evolution*, *6*(9), 1390–1397. https://doi.org/10.1038/s41559-022-01820-0

Qiu, Y. (2024). *showtext: Using fonts more easily in r graphs* (version 0.9-7) [R package]. https://github.com/yixuan/showtext

R Core Team. (2024). *R: A language and environment for statistical computing* (version 4.4.0). R Foundation for Statistical Computing. https://www.R-project.org/

Railsback, S. F., Gallagher, C. A., Grimm, V., McCary, M. A., & Harvey, B. C. (2025). Empirical ecology to support mechanistic modelling: Different objectives, better approaches and unique benefits. *Methods in Ecology and Evolution*, *16*(8), 1564–1573. https://doi.org/10.1111/2041-210X.70083

Regan, H. M., Ben-Haim, Y., Langford, B., Wilson, G., William, Lundberg, P., Andelman, S. J., & Burgman, M. (2005). Robust decision-making under severe uncertainty for conservation management. *Ecological Applications*, *15*(4), 1471–1477. https://doi.org/10.1890/03-5419

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705. https://doi.org/10.1126/science.1197962

Rijnhart, J. J. M., Twisk, J. W. R., Deeg, D. J. H., & Heymans, M. W. (2021). Assessing the robustness of mediation analysis results using multiverse analysis. *Prevention Science: The Official Journal of the Society for Prevention Research*, *23*(5), 821–831. https://doi.org/10.1007/s11121-021-01280-1

Riordan-Short, E., Pither, R., & Pither, J. (2023). Four steps to strengthen connectivity modeling. *Ecography*, *2023*(11), e06766. https://doi.org/10.1111/ecog.06766

Risbey, J., van der Sluijs, J., Kloprogge, P., Ravetz, J., Funtowicz, S., & Quintana, S. C. (2005). Application of a checklist for quality assistance in environmental modelling to an energy model. *Environmental Modeling & Assessment*, *10*(1), 63–79. https://doi.org/10.1007/s10666-004-4267-z

Roche, D. G., O'Dea, R. E., Kerr, K. A., Rytwinski, T., Schuster, R., Nguyen, V. M., Young, N., Bennett, J. R., & Cooke, S. J. (2021). Closing the knowledge-action gap in conservation with open science. *Conservation Biology*, *36*(3), e13835. https://doi.org/10.1111/cobi.13835

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 1. https://doi.org/10.5334/labphon.147

Rohrer, J. M., Smith, G. D., & Munafò, M. (2025). What can be learned when multiple analysts arrive at different estimates. *European Journal of Epidemiology*, *40*(5), 493–495. https://doi.org/10.1007/s10654-025-01249-2

Rosenberg, M. S. (2013). Chapter 9. Moment and least-squares based approaches to metaanalytic inference. In J. Koricheva, J. Gurevitch, & K. Mengersen (Eds.), *Handbook of meta-analysis in ecology and evolution* (pp. 108–124). Princeton University Press. https://doi.org/10.1515/9781400846184-011

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, *15*(1), 1829. https://doi.org/10.1038/s41467-024-46150-w

Rounsevell, M. D. A., Arneth, A., Brown, C., Cheung, W. W. L., Gimenez, O., Holman, I., Leadley, P., Luján, C., Mahevas, S., Maréchaux, I., Pélissier, R., Verburg, P. H., Vieilledent, G., Wintle, B. A., & Shin, Y.-J. (2021). Identifying uncertainties in scenarios and models of socio-ecological systems in support of decision-making. *One Earth*, *4*(7), 967–985. https://doi.org/10.1016/j.oneear.2021.06.003

Royle, N. J., Hartley, I. R., Owens, I. P. F., & Parker, G. A. (1999). Sibling competition and the evolution of growth rates in birds. *Proceedings of the Royal Society B: Biological Sciences*, *266*(1422), 923–932. https://doi.org/10.1098/rspb.1999.0725

Rykiel Jr., E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling*, *90*(3), 229–244. https://doi.org/10.1016/0304-3800(95)00152-2

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29–39.

Sánchez-Tójar, A., Bezine, A., Purgar, M., & Čulina, A. (2025). Code-sharing policies are associated with increased reproducibility potential of ecological findings. *Peer Community Journal*, *5*(e37). https://doi.org/10.24072/pcjournal.541

Sarma, A., Kale, A., Moon, M. J., Taback, N., Chevalier, F., Hullman, J., & Kay, M. (2023). Multiverse: Multiplexing alternative data analyses in R notebooks. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3544548.3580726

Scheinin, I., Kalimeri, M., Jagerroos, V., Parkkinen, J., Tikkanen, E., Würtz, P., & Kangas, A. (2020). *ggforestplot: Forestplots of measures of effects and their confidence intervals* (version 0.1.0, commit 547617e63fa481a5f28ffc56c07d46be4af688b2) [R package]. https://github.com/NightingaleHealth/ggforestplot

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2024). *GGally: Extension to "ggplot2"* (version 2.2.1) [R package]. https://ggobi.github.io/ggally/

Schmolke, A., Thorbek, P., Deangelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: A strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479–486. https://doi.org/10.1016/j.tree.2010.05.001

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S. D., Martínez-López, J., & Vermeiren, P. (2019). How to make ecological models useful for environmental management. *Ecological Modelling*, *411*, 108784. https://doi.org/10.1016/j.ecolmodel.2019.108784

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., Aert, R. C. M. van, van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., … Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Seaman, C. S., & Weber, R. (2015). Undisclosed flexibility in computing and reporting structural equation models in communication science. *Communication Methods and Measures*, *9*(4), 208–232. https://doi.org/10.1080/19312458.2015.1096329

Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., & Kowald, D. (2023). *Reproducibility in machine learning-driven research* [Preprint]. *arXiv:2307.10320*. http://arxiv.org/abs/2307.10320

Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E. S. A., & Nakagawa, S. (2016). Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications. *Ecology*, *97*(12), 3293–3299. https://doi.org/10.1002/ecy.1591

Seppelt, R., Beckmann, M., Václavík, T., & Volk, M. (2018). The art of scientific performance. *Trends in Ecology & Evolution*, *33*(11), 805–809. https://doi.org/10.1016/j.tree.2018.08.003

Sequeira, A. M. M., O'Toole, M., Keates, T. R., McDonnell, L. H., Braun, C. D., Hoenner, X., Jaine, F. R. A., Jonsen, I. D., Newman, P., Pye, J., Bograd, S. J., Hays, G. C., Hazen, E. L., Holland, M., Tsontos, V. M., Blight, C., Cagnacci, F., Davidson, S. C., Dettki, H., … Weise, M. (2021). A standardisation framework for bio-logging data to advance ecological research and conservation. *Methods in Ecology and Evolution*, *12*(6), 996–1007. https://doi.org/10.1111/2041-210x.13593

Shmueli, G. (2010). To explain or to predict. *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shoemaker, K. T., & Loope, K. J. (2025). We need better ways to re-evaluate conservation policies when they're founded on flawed research. *Proceedings of the National Academy of Sciences*, *122*(19), e2426166122. https://doi.org/10.1073/pnas.2426166122

Siegel, K. R., Kaur, M., Grigal, A. C., Metzler, R. A., & Dickinson, G. H. (2022). Meta-analysis suggests negative, but pCO2-specific, effects of ocean acidification on the structural and functional properties of crustacean biomaterials. *Ecology and Evolution*, *12*(6), e8922. https://doi.org/10.1002/ece3.8922

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, *1*(3), 27. https://doi.org/10.21105/joss.00037

Silk, M. J., Harrison, X. A., & Hodgson, D. J. (2020). Perils and pitfalls of mixed-effects regression models in biology. *PeerJ*, *8*, e9522. https://doi.org/10.7717/peerj.9522

Sillero, N. (2011). What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, *222*(8), 1343–1346. https://doi.org/10.1016/j.ecolmodel.2011.01.018

Simmonds, E. G., Adjei, K. P., Cretois, B., Dickel, L., González-Gil, R., Laverick, J. H., Mandeville, C. P., Mandeville, E. G., Ovaskainen, O., Sicacha-Parada, J., Skarstein, E. S., & O'Hara, B. (2024). Recommendations for quantitative uncertainty consideration in ecology and evolution. *Trends in Ecology & Evolution*, *39*(4), 328–337. https://doi.org/10.1016/j.tree.2023.10.012

Simmonds, M. B., Riley, W. J., Agarwal, D. A., Chen, X., Cholia, S., Crystal-Ornelas, R., Coon, E. T., Dwivedi, D., Hendrix, V. C., Huang, M., Jan, A., Kakalia, Z., Kumar, J., Koven, C. D., Li, L., Melara, M., Ramakrishnan, L., Ricciuto, D. M., Walker, A. P., … Varadharajan, C. (2022). Guidelines for publicly archiving terrestrial model data to enhance usability, intercomparison, and synthesis. *Data Science Journal*, *21*(1), 3. https://doi.org/10.5334/dsj-2022-003

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/09567976114176

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(1215), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Sjoberg, D. D., Whiting, K., Curry, M., Lavery, J. A., & Larmarange, J. (2021). Reproducible summary tables with the gtsummary package. *The R Journal*, *13*, 570–580. https://doi.org/10.32614/RJ-2021-053

Slowikowski, K. (2024). *ggrepel: Automatically position non-overlapping text labels with "ggplot2"* (version 0.9.5) [R package]. https://ggrepel.slowkow.com/

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*(4), 207–218. https://doi.org/10.1027/1864-9335/a000425

Srivastava, S. (2018). Sound inference in complicated research: A multi-strategy approach [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/bwr48

Stanton-Geddes, J., de Freitas, C. G., & de Sales, D. C. (2014). In defense of p values: Comment on the statistical methods actually used by ecologists. *Ecology*, *95*(3), 637–642. https://doi.org/10.1890/13-1156.1

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stephens, P. A., Buskirk, S. W., & del Rio, C. M. (2007). Inference in ecology and evolution. *Trends in Ecology & Evolution*, *22*(4), 192–197. https://doi.org/10.1016/j.tree.2006.12.003

Stewart Lowndes, J. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(0160). https://doi.org/10.1038/s41559-017-0160

Stock, A., Gregr, E. J., & Chan, K. M. A. (2023). Data leakage jeopardizes ecological applications of machine learning. *Nature Ecology & Evolution*, *7*(11), 1743–1745. https://doi.org/10.1038/s41559-023-02162-1

Stoudt, S., Vásquez, V. N., & Martinez, C. C. (2021). Principles for data analysis workflows. *PLOS Computational Biology*, *17*(3), e1008770. https://doi.org/10.1371/journal.pcbi.1008770

Sutherland, W. J., & Wordley, C. F. R. (2017). Evidence complacency hampers conservation. *Nature Ecology & Evolution*, *1*(9), 1–2. https://doi.org/10.1038/s41559-017-0244-1

Szetey, K., Ward, D., Chakori, S., Douglas, D., Fulton, E. A., Grigg, N., Ligtermoet, E., Munera-Roldan, C., Onyango, E., Stirling, E., & Subramaniam, R. (2025). Applying an ethical lens for more responsible modelling practice. *Socio-Environmental Systems Modelling*, *7*, 18753. https://doi.org/10.18174/sesmo.18753

Taylor, J. W., & Taylor, K. S. (2023). Combining probabilistic forecasts of COVID-19 mortality in the united states. *European Journal of Operational Research*, *304*(1), 25–41. https://doi.org/10.1016/j.ejor.2021.06.044

Terry, J., Christopher, D., & Armitage, D. W. (2024). Widespread analytical pitfalls in empirical coexistence studies and a checklist for improving their statistical robustness. *Methods in Ecology and Evolution*, *15*(4), 594–611. https://doi.org/10.1111/2041-210X.14227

Thibault, R. T., Kovacs, M., Hardwicke, T. E., Sarafoglou, A., Ioannidis, J. P. A., & Munafò, M. R. (2023). Reducing bias in secondary data analysis via an explore and confirm analysis workflow (ECAW): A proposal and survey of observational researchers. *Royal Society Open Science*, *10*(10), 230568. https://doi.org/10.1098/rsos.230568

Thrall, P. H., Chase, J., Drake, J., Espuno, N., Hello, S., Ezenwa, V., Han, B., Mori, A., & Muller-Landau, H. (2023). From raw data to publication: Introducing data editing at ecology letters. *Ecology Letters*, *26*(6), 829–830. https://doi.org/10.1111/ele.14210

Tierney, N., & Cook, D. (2023). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, *105*(7), 1–31. https://doi.org/10.18637/jss.v105.i07

Tincani, M., & Travers, J. (2019). Replication Research, Publication Bias, and Applied Behavior Analysis. *Perspectives on Behavior Science*, *42*(1), 59–75. https://doi.org/10.1007/s40614-019-00191-5

Todman, L. C., Bush, A., & Hood, A. S. C. (2023). 'Small data' for big insights in ecology. *Trends in Ecology & Evolution*, *38*(7), 615–622. https://doi.org/10.1016/j.tree.2023.01.015

Touchon, J. C., & McCoy, M. W. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere*, *7*(8), e01394. https://doi.org/10.1002/ecs2.1394

Travers, H., Selinske, M., Nuno, A., Serban, A., Mancini, F., Barychka, T., Bush, E., Rasolofoson, R. A., Watson, J. E. M., & Milner-Gulland, E. J. (2019). A manifesto for predictive conservation. *Biological Conservation*, *237*, 12–18. https://doi.org/10.1016/j.biocon.2019.05.059

Traylor, W. (2024). Model-based experiments as epistemic evidence in paleoecology. *Ecological Modelling*, *498*, 110895. https://doi.org/10.1016/j.ecolmodel.2024.110895

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336. https://doi.org/10.1002/ecy.3336

Upmeier zu Belzen, A., Engelschalt, P., & Krüger, D. (2021). Modeling as scientific reasoning—the role of abductive reasoning for modeling competence. *Education Sciences*, *11*(9), 495. https://doi.org/10.3390/educsci11090495

Ushey, K., & Wickham, H. (2024). *Renv: Project environments*. https://doi.org/10.32614/CRAN.package.renv R package version 1.0.7

Ushey, K., & Wickham, H. (2026). *renv: Project environments*. https://doi.org/10.32614/CRAN.package.renv R package version 1.1.7

van den Akker, O. R., Bakker, M., Van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. M. (2024). The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. *Psychological Methods*. https://doi.org/10.1037/met0000687

van den Brand, T. (2024). *ggh4x: Hacks for "ggplot2"* (version 0.2.8) [R package]. https://github.com/teunbrand/ggh4x

Van Der Sluijs, J. P. (2002). A way out of the credibility crisis of models used in integrated environmental assessment. *Futures*, *34*(2), 133–146. https://doi.org/10.1016/S0016-3287(01)00051-9

Van Eijk, N. L., Jiao, H., & Peters, G.-J. Y. (2023). *Making Preregistration Accessible: An R Package to Make Machine-Readable Preregistrations and Create New Preregistration Forms*. https://doi.org/10.31234/osf.io/qfjy7

Vander Werf, E. (1992). Lack's clutch size hypothesis: An examination of the evidence using meta-analysis. *Ecology*, *73*(5), 1699–1705. https://doi.org/10.2307/1940021

Ver Hoef, J. M. (2012). Who invented the delta method? *The American Statistician*, *66*(2), 124–127. https://doi.org/10.1080/00031305.2012.687494

Verhulst, S., Holveck, M. J., & Riebel, K. (2006). Long-term effects of manipulated natal brood size on metabolic rate in zebra finches. *Biology Letters*, *2*(3), 478–480. https://doi.org/10.1098/rsbl.2006.0496

Vesk, P. A., Morris, W. K., McCallum, W., Apted, R., & Miles, C. (2016). Processes of woodland eucalypt regeneration: Lessons from the bush returns trial. *Proceedings of the Royal Society of Victoria*, *128*(1), 54–63. https://doi.org/10.1071/RS16005

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how. *Zeitschrift Für Psychologie*, *227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Ware, J. J., & Munafò, M. R. (2015). Significance chasing in research practice: Causes, consequences and possible solutions. *Addiction*, *110*(1), 4–8. https://doi.org/10.1111/add.12673

Weiskopf, S. R., Harmáčková, Z. V., Johnson, C. G., Londoño-Murcia, M. C., Miller, B. W., Myers, B. J. E., Pereira, L., Arce-Plata, M. I., Blanchard, J. L., Ferrier, S., Fulton, E. A., Harfoot, M., Isbell, F., Johnson, J. A., Mori, A. S., Weng, E., & Rosa, I. M. D. (2022). Increasing the uptake of ecological model results in policy decisions to improve biodiversity outcomes. *Environmental Modelling & Software*, *149*, 105318. https://doi.org/10.1016/j.envsoft.2022.105318

Westoby, M., Falster, D. S., & Schrader, J. (2021). Motivating data contributions via a distinct career currency. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1946), 20202830. https://doi.org/10.1098/rspb.2020.2830

White, C. R., & Marshall, D. J. (2019). Should we care if models are phenomenological or mechanistic. *Trends in Ecology & Evolution*, *34*(4), 276–278. https://doi.org/10.1016/j.tree.2019.01.006

White, E. P., Yenni, G. M., Taylor, S. D., Christensen, E. M., Bledsoe, E. K., Simonis, J. L., & Ernest, S. K. M. (2019). Developing an automated iterative near-term forecasting system for an ecological study. *Methods in Ecology and Evolution*, *10*(3), 332–344. https://doi.org/10.1111/2041-210x.13104

White, N., Parsons, R., Collins, G., & Barnett, A. (2023). Evidence of questionable research practices in clinical prediction models. *BMC Medicine*, *21*(1). https://doi.org/10.1186/s12916-023-03048-6

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and

reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832). https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2011). Testthat: Get started with testing. *The R Journal*, *3*, 5–10. https://journal.r-project.org/articles/RJ-2011-002/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2023). *R packages (2e)*. O'Reilly Media, Inc. https://r-pkgs.org

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data* (2nd ed.). O'Reilly Media. https://r4ds.hadley.nz/

Wickham, H., Hesselberth, J., Salmon, M., Roy, O., & Brüggemann, S. (2024). *Pkgdown: Make static HTML documentation for a package* (version 2.1.1) [R package]. https://pkgdown.r-lib.org/

Wickham, H., Hester, J., Chang, W., & Bryan, J. (2022). *devtools: Tools to make developing r packages easier* (version 2.4.5) [R package].

Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *scales: Scale functions for visualization* (version 1.3.0) [R package]. https://scales.r-lib.org

Wickham, H., Vaughan, D., & Girlich, M. (2025). *Tidyr: Tidy messy data*. https://doi.org/10.32614/CRAN.package.tidyr R package version 1.3.2

Wigboldus, D. H. J., & Dotsch, R. (2015). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*, *81*(1), 1–6. https://doi.org/10.1007/s11336-015-9445-1

Wilke, C. O. (2024). *cowplot: Streamlined plot theme and plot annotations for "ggplot2"* (version 1.1.3) [R package]. https://wilkelab.org/cowplot/

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R.,

… Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Williams, C., Yang, Y., Lagisz, M., Morrison, K., Ricolfi, L., Warton, D. I., & Nakagawa, S. (2024). Transparent reporting items for simulation studies evaluating statistical methods: Foundations for reproducibility and reliability. *Methods in Ecology and Evolution*, *15*(11), 1926–1939. https://doi.org/10.1111/2041-210X.14415

Williams, C., Yang, Y., Warton, D. I., & Nakagawa, S. (2025). Modelling approaches for meta-analyses with dependent effect sizes in ecology and evolution: A simulation study. *Methods in Ecology and Evolution*, *16*(10), 2362–2379. https://doi.org/10.1111/2041-210X.70156

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231213802

Woo, S. E., O'Boyle, E. H., & Spector, P. E. (2017). Best practices in developing, conducting, and evaluating inductive research. *Human Resource Management Review*, *27*(2), 255–264. https://doi.org/10.1016/j.hrmr.2016.08.004

Wood, C. M., Loman, Z. G., McKinney, S. T., & Loftin, C. S. (2020). Testing prediction accuracy in short-term ecological studies. *Basic and Applied Ecology*, *43*, 77–85. https://doi.org/https://doi.org/10.1016/j.baae.2020.01.003

Xie, Y. (2024a). *knitr: A general-purpose package for dynamic report generation in r* (version 1.46) [R package]. https://yihui.org/knitr

Xie, Y. (2024b). *xfun: Supporting functions for packages maintained by "Yihui Xie"* (version 0.44) [R package]. https://github.com/yihui/xfun

Yanco, S. W. et al. (2020). A modern method of multiple working hypotheses to improve inference in ecology. *Royal Society Open Science*, *7*(6), 200231.

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., & Nakagawa, S. (2022). Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global Change Biology*, *28*(3), 969–989. https://doi.org/10.1111/gcb.15972

Yang, Y., Sánchez-Tójar, A., O'Dea, R. E., Noble, D. W. A., Koricheva, J., Jennions, M. D., Parker, T. H., Lagisz, M., & Nakagawa, S. (2023). Publication bias impacts on effect size, statistical power, and magnitude (type m) and sign (type s) errors in ecology

and evolutionary biology. *BMC Biology*, *21*(1), 71. https://doi.org/10.1186/s12915-022-01485-y

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*. https://doi.org/10.1017/s0140525x20001685

Yates, K., Bouchet, P., Caley, M., Mengersen, K., Randin, C., Parnell, S., Fielding, A., Bamford, A., Ban, S., Barbosa, A., Dormann, C., Elith, J., Embling, C., Ervin, G., Fisher, R., Gould, S., Graf, R., Gregr, E., Halpin, P., … Sequeira, A. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Yenni, G., Christensen, E., Bledsoe, E., Supp, S., Diaz, R., White, E. P., & Ernest, S. (2019). Developing a modern data workflow for regularly updated data. *PLOS Biology*, *17*(1), e3000125. https://doi.org/10.1371/journal.pbio.3000125

Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, *12*(11), 2117–2128. https://doi.org/10.1111/2041-210X.13686

Zagrodzka, Z., Johnson, T. F., & Beckerman, A. P. (2024). Accelerating the open research agenda to solve global challenges. *Ecology and Evolution*. https://doi.org/10.1002/ece3.10887

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2020). colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software*, *96*(1), 1–49. https://doi.org/10.18637/jss.v096.i01

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277. https://doi.org/10.1111/ecog.04960

Zvereva, E. L., & Kozlov, M. V. (2021). Biases in ecological research: Attitudes of scientists and ways of control. *Scientific Reports*, *11*(226). https://doi.org/10.1038/s41598-020-80677-4

A Supporting Information for the "Same Data, Different Analysts" Study

Supporting Information for Chapter 2. The appendices described below correspond to Supplementary Information 1 in Gould et al. (2025), Chapter 2, where: SM A is Appendix A.1, SM B is Appendix A.2, SM C is Appendix A.3 and SM D is Section A.4. Appendix 1 and 2 of Gould et al. (2025) are reproduced in Appendix A.5.

A.1 Summarising Variation Among Analysis Specifications

A.1.1 Summary Statistics

Number of Analyses of Different Types

As described in the summary statistics section of the manuscript, 63 teams submitted 131 Z_r model estimates and 43 teams submitted 64 y_i predictions for the blue tit dataset. Similarly, 40 submitted 79 Z_r model estimates and 14 teams submitted 24 y_i predictions for the *Eucalytpus* dataset. The majority of the blue tit analyses specified normal error distributions and were non-Bayesian mixed effects models. Analyses of the *Eucalyptus* dataset rarely specified normal error distributions, likely because the response variable was in the form of counts. Mixed effects models were also common for *Eucalytpus* analyses (Table A.1).

Model Composition

The composition of models varied substantially (Table A.2) in regards to the number of fixed and random effects, interaction terms and the number of data points used. For the blue tit dataset, models used up to 19 fixed effects, 12 random effects, and 10 interaction terms and had sample sizes ranging from 76 to 3720. For the *Eucalyptus* dataset models

Table A.1: *Summary of the number of analysis teams, total analyses, models with normal error distributions, mixed effects models, and models developed with Bayesian statistical methods for effect size analyses only (Z_r) and out-of-sample prediction only (y_i).*

	No. Analyses	No. Teams	Normal Distribution	Mixed Effect	Bayesian
blue tit					
Z_r	131	63	124	128	10
y_i	64	43	59	63	10
Eucalyptus					
Z_r	79	40	15	62	5
y_i	24	14	1	16	3

had up to 13 fixed effects, 4 random effects, 5 interaction terms and sample sizes ranging from 18 to 351.

Choice of Variables

The choice of variables also differed substantially among analyses (Table A.3) and some analysts constructed new variables that transformed or aggregated one or more existing variables. The blue tit dataset had a total of 52 candidate variables. These variables were included in a mean of 20.5 Z_r analyses (range 0- 100). The *Eucalyptus* dataset had a total of 59 candidate variables. The variables in the *Eucalyptus* dataset were included in a mean of 8.92 Z_r analyses (range 0-55).

Table A.2: *Mean, standard deviation and range of number of fixed and random variables, interaction terms used in models and analysis sample size (N). Repeated for effect-size analyses only (Z_r) and out-of-sample predictions only (y_i).*

	Mean		SD		Min		Max	
	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*
Fixed								
Z_r	5.20	5.01	2.92	3.83	1	1	19	13
y_i	4.78	4.67	2.35	3.45	1	1	10	13
Random								
Z_r	3.53	1.41	2.08	1.09	0	0	10	4
y_i	4.42	0.96	2.78	0.81	1	0	12	3
Interactions								
Z_r	0.44	0.16	1.11	0.65	0	0	10	5
y_i	0.28	0.17	0.63	0.48	0	0	3	2
N								
Z_r	2611.09	298.43	937.48	106.25	76	18	3720	351
y_i	2816.71	325.55	773.21	64.17	396	90	3720	350

Table A.3: *Mean, SE, minimum and maximum number of analyses in which each variable was used, for effect size analyses only (Z_r), out-of-sample predictions only (y_i), using the full dataset.*

	Mean		SD		Min		Max	
	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*	blue tit	*Eucalyptus*
Z_r	20.5	8.92	27	12.28	0	0	100	55
y_i	10.79	2.2	13.87	3.7	0	0	52	17

A.1.2 Effect Size Specification Analysis

We used a specification curve (Simonsohn et al., 2015) to look for relationships between Z_r values and several modeling decisions, including the choice of independent and dependent variable, transformation of the dependent variable, and other features of the models that produced those Z_r values (Figure A.1, Figure A.2). Each effect can be matched to the model features that produced it by following a vertical line down the figure.

Blue Tit

We observed few clear trends in the blue tit specification curve (Figure A.1). The clearest trend was for the independent variable 'contrast: reduced broods vs. unmanipulated broods' to produce weak or even positive relationships, but never strongly negative relationships. The biological interpretation of this pattern is that nestlings in reduced broods averaged similar growth to nestlings in unmanipulated broods, and sometimes the nestlings in reduced broods even grew less than the nestlings in unmanipulated broods. Therefore, it may be that competition limits nestling growth primarily when the number of nestlings exceeds the clutch size produced by the parents, and not in unmanipulated broods. The other relatively clear trend was that the strongest negative relationships were never based on the independent variable 'contrast: unmanipulated broods vs. enlarged broods'. These observations demonstrate the potential value of specification curves.

Eucalyptus

In the *Eucalyptus* specification curve, there are no strong trends (Figure A.2). It is, perhaps, the case that choosing the dependent variable 'count of seedlings 0-0.5m high' corresponds to more positive results and choosing 'count of all *Eucalytpus* seedlings' might find more negative results. Choosing the independent variable 'sum of all grass types (with or without non-grass graminoids)' might be associated with more results close to zero consistent with the absence of an effect.

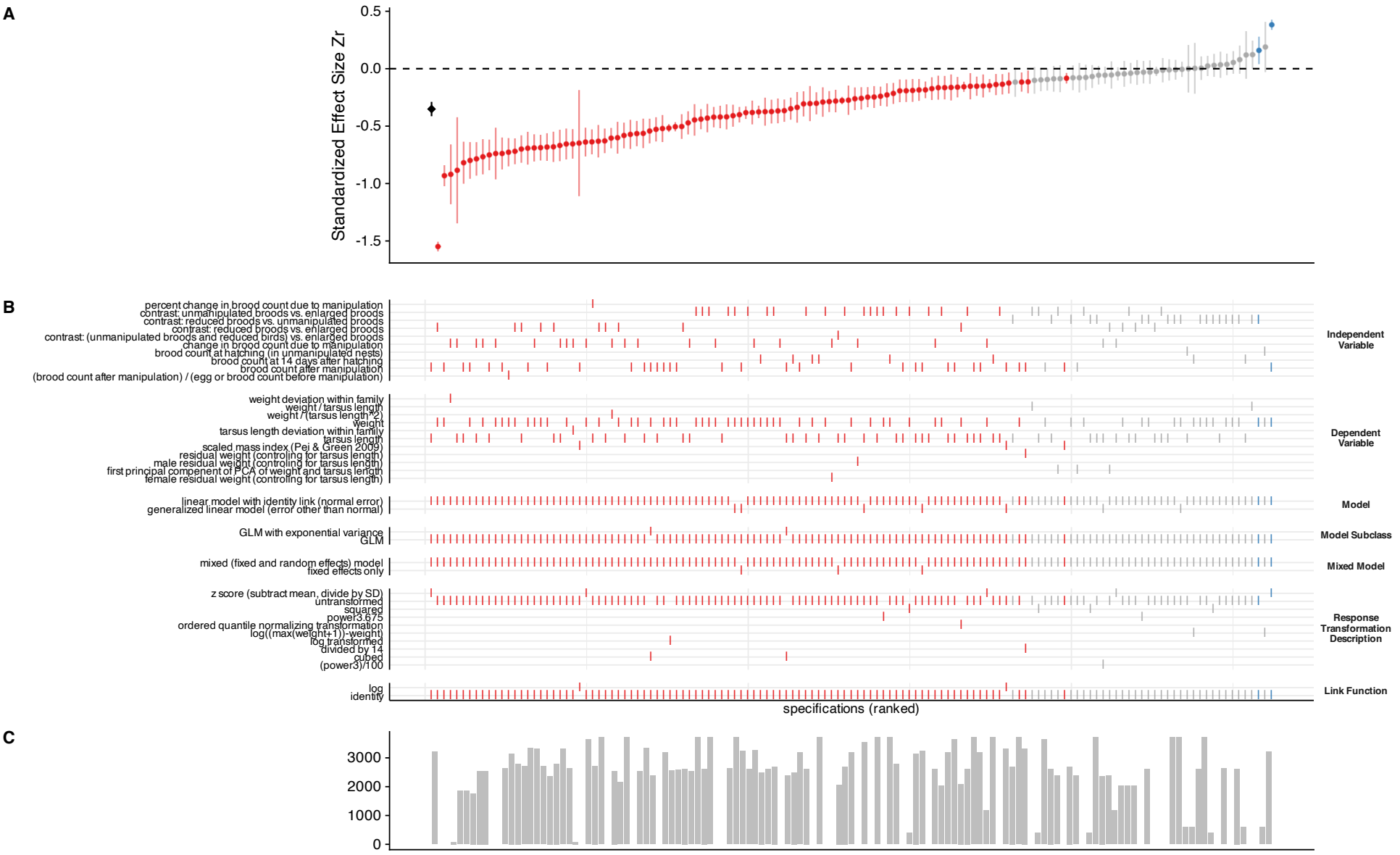


Figure A.1: ***A.*** *Forest plot for blue tit analyses: standardised effect-sizes (circles) and their 95% confidence intervals are displayed for each analysis included in the meta-analysis model. The meta-analytic mean effect-size is denoted by a black diamond, with error bars also representing the 95% confidence interval. The dashed black line demarcates effect sizes of 0, whereby no effect of the test variable on the response variable is found. Blue points where Z_r and its associated confidence intervals are greater than 0 indicate analyses that found a negative effect of sibling number on nestling growth. Gray coloured points have confidence intervals crossing 0, indicating no relationship between the test and response variable. Red points indicate the analysis found a positive relationship between sibling number and nestling growth.* ***B.*** *Analysis specification plot: for each analysis plotted in A, the corresponding combination of analysis decisions is plotted. Each decision and its alternative choices is grouped into its own facet, with the decision point described on the right of the panel, and the option shown on the left. Lines indicate the option chosen used in the corresponding point in plot A.* ***C.*** *Sample sizes of each analysis. Note that empty bars indicate analyst did not report sample size and sample size could not be derived by lead team.*

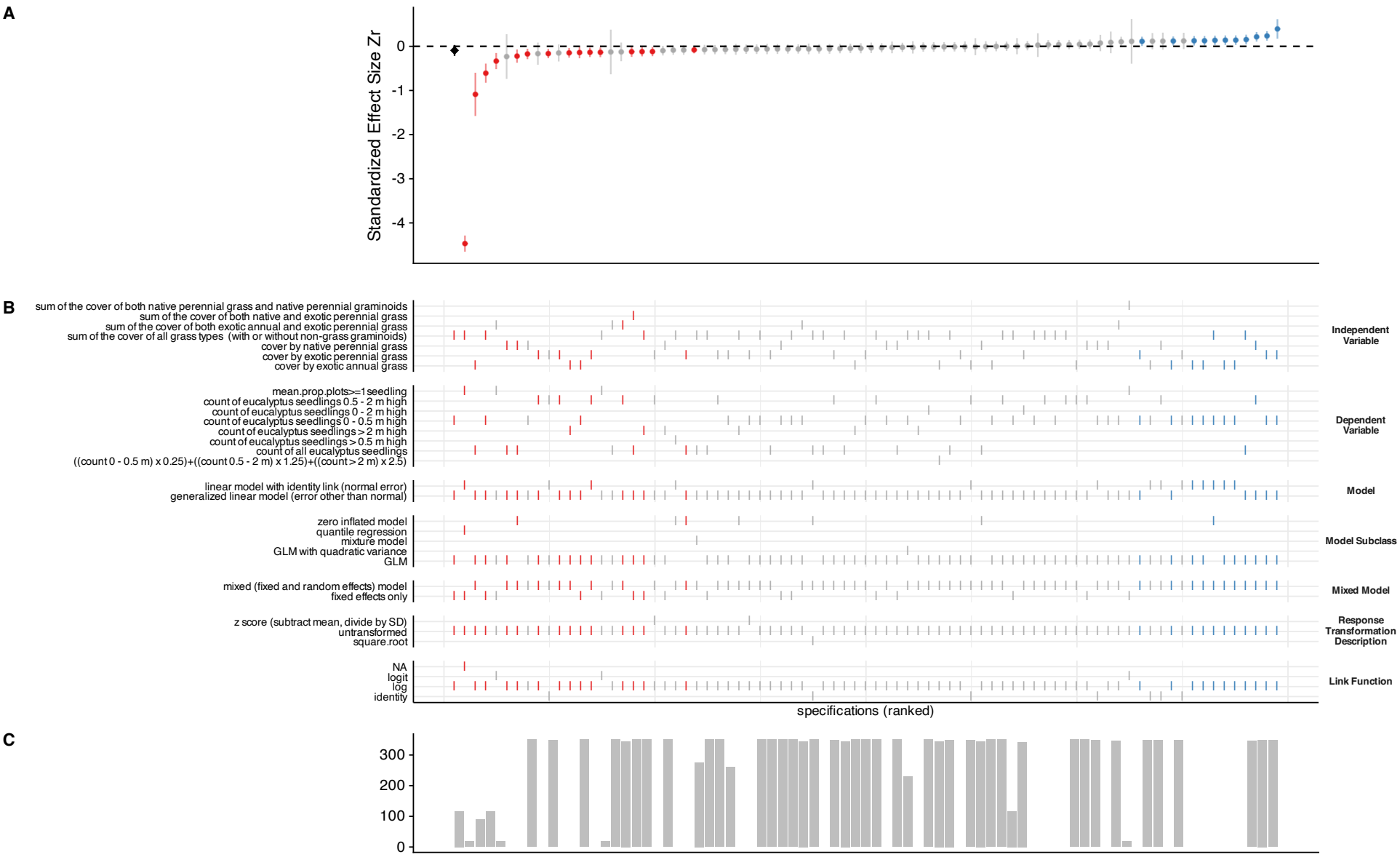


Figure A.2: ***A.*** *Forest plot for* Eucalyptus *analyses: standardised effect-sizes (circles) and their 95% confidence intervals are displayed for each analysis included in the meta-analysis model. The meta-analytic mean effect-size is denoted by a black diamond, with error bars also representing the 95% confidence interval. The dashed black line demarcates effect sizes of 0, whereby no effect of the test variable on the response variable is found. Blue points where* Z_r *and its associated confidence intervals are greater than 0 indicate analyses that found a positive relationship of grass cover on* Eucalyptus *seedling success. Gray coloured points have confidence intervals crossing 0, indicating no relationship between the test and response variable. Red points indicate the analysis found a negative relationship between grass cover and* Eucalyptus seedling success. ***B.*** *Analysis specification plot: for each analysis plotted in A, the corresponding combination of analysis decisions is plotted. Each decision and its alternative choices is grouped into its own facet, with the decision point described on the right of the panel, and the option shown on the left. Lines indicate the option chosen used in the corresponding point in plot A.* ***C.*** *Sample sizes of each analysis. Note that empty bars indicate analyst did not report sample size and sample size could not be derived by lead team.*

A.2 Effect Size Analysis

A.2.1 Meta-analysis

Effect Sizes Z_r

Effect of Categorical Review Rating The figures below (Figure A.3, Figure A.4) hows the fixed effect of categorical review rating on deviation from the meta-analytic mean. There is very little difference in deviation for analyses in any of the review categories. It is worth noting that each analysis features multiple times in these figures corresponding to the multiple reviewers that provided ratings.

Post-hoc Analysis: Exploring the effect of removing analyses with poor peer-review ratings on heterogeneity The forest plots in Figure A.6 compare the distributions of Z_r effects from our full set of analyses with the distributions of Z_r effects from our post-hoc analyses, which removed either analyses that were reviewed at least once as being 'unpublishable', and analyses that were reviewed at least once as being 'unpublishable' or requiring 'major revisions'. Removing these analyses from the blue tit data had little impact

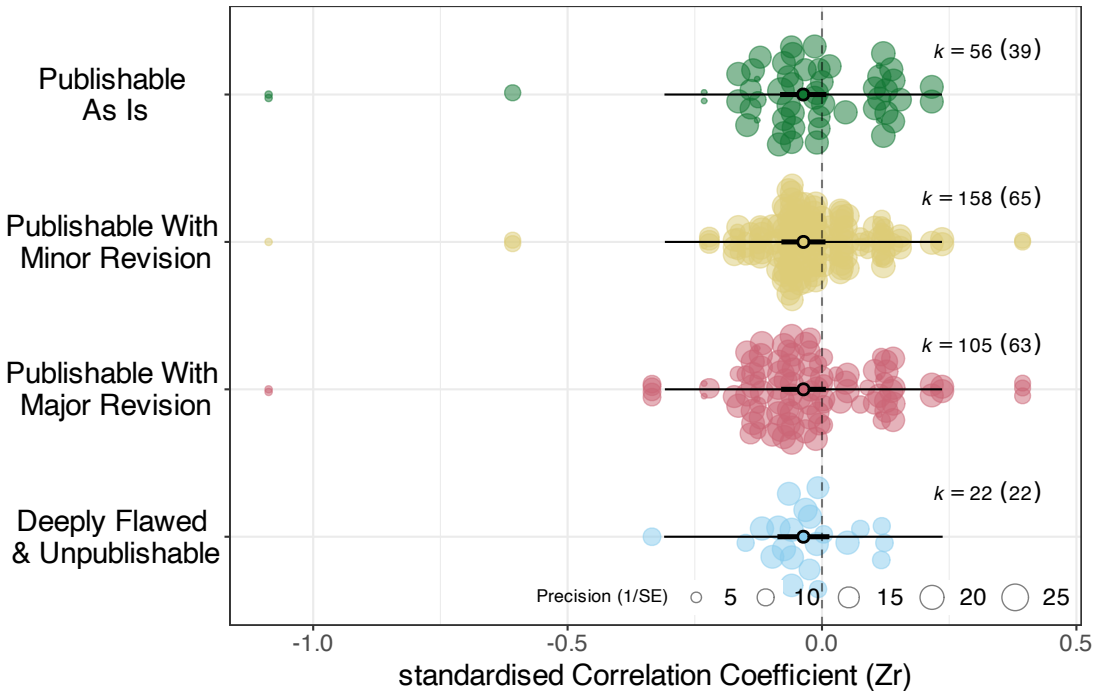


Figure A.3: *Orchard plot of meta-analytic model fitted to all* Eucalyptus *analyses with a fixed effect for categorical peer-review ratings, and random effects for analyst ID and reviewer ID. Black circles denote coefficient mean for each categorical publishability rating. Thick error bars represent 95% confidence intervals and whiskers indicate 95% prediction intervals. Effect sizes are represented by circles and their size corresponds to the precision of the estimate.*

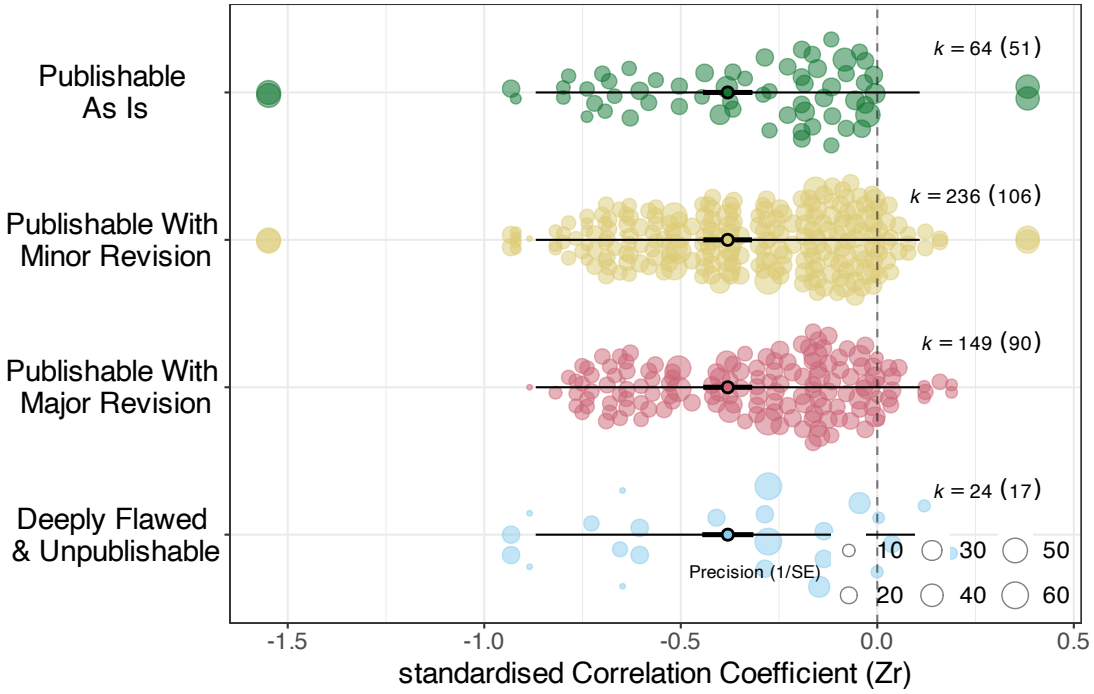


Figure A.4: *Orchard plot of meta-analytic model fitted to all blue tit analyses with a fixed effect for categorical peer-review ratings, and random effects for analyst ID and reviewer ID. Black circles denote coefficient mean for each categorical publishability rating. Thick error bars represent 95% confidence intervals and whiskers indicate 95% prediction intervals. Effect sizes are represented by circles and their size corresponds to the precision of the estimate.*

on the overall distribution of the results. When 'unpublishable' analyses of the *Eucalyptus* dataset were removed, the extreme outlier 'Brooklyn-2-2-1' was also removed, resulting in a substantial difference to the amount of observed deviation from the meta-analytic mean.

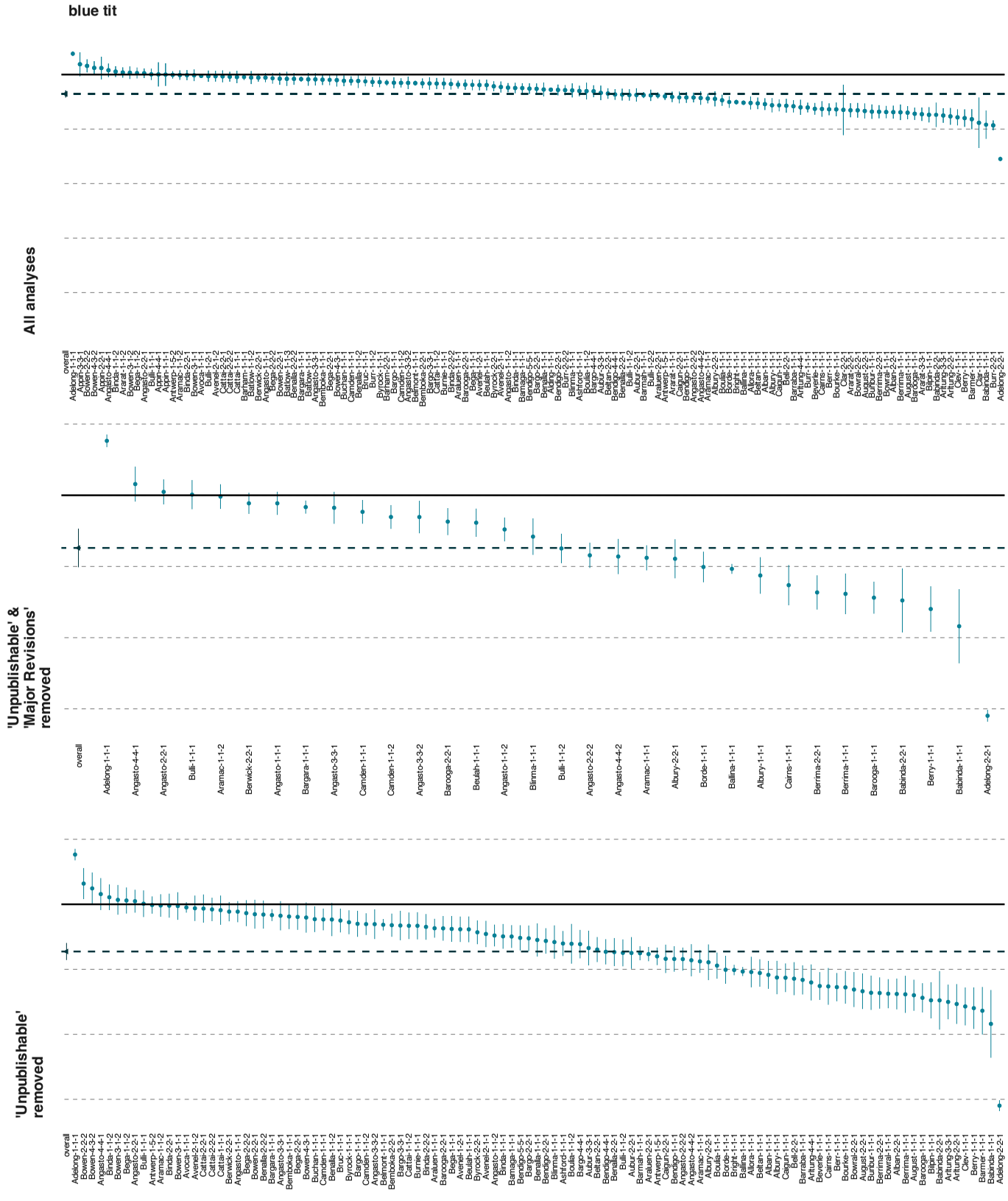

blue tit
All analyses
'Unpublishable' & 'Major Revisions' removed
'Unpublishable' removed

Figure A.6

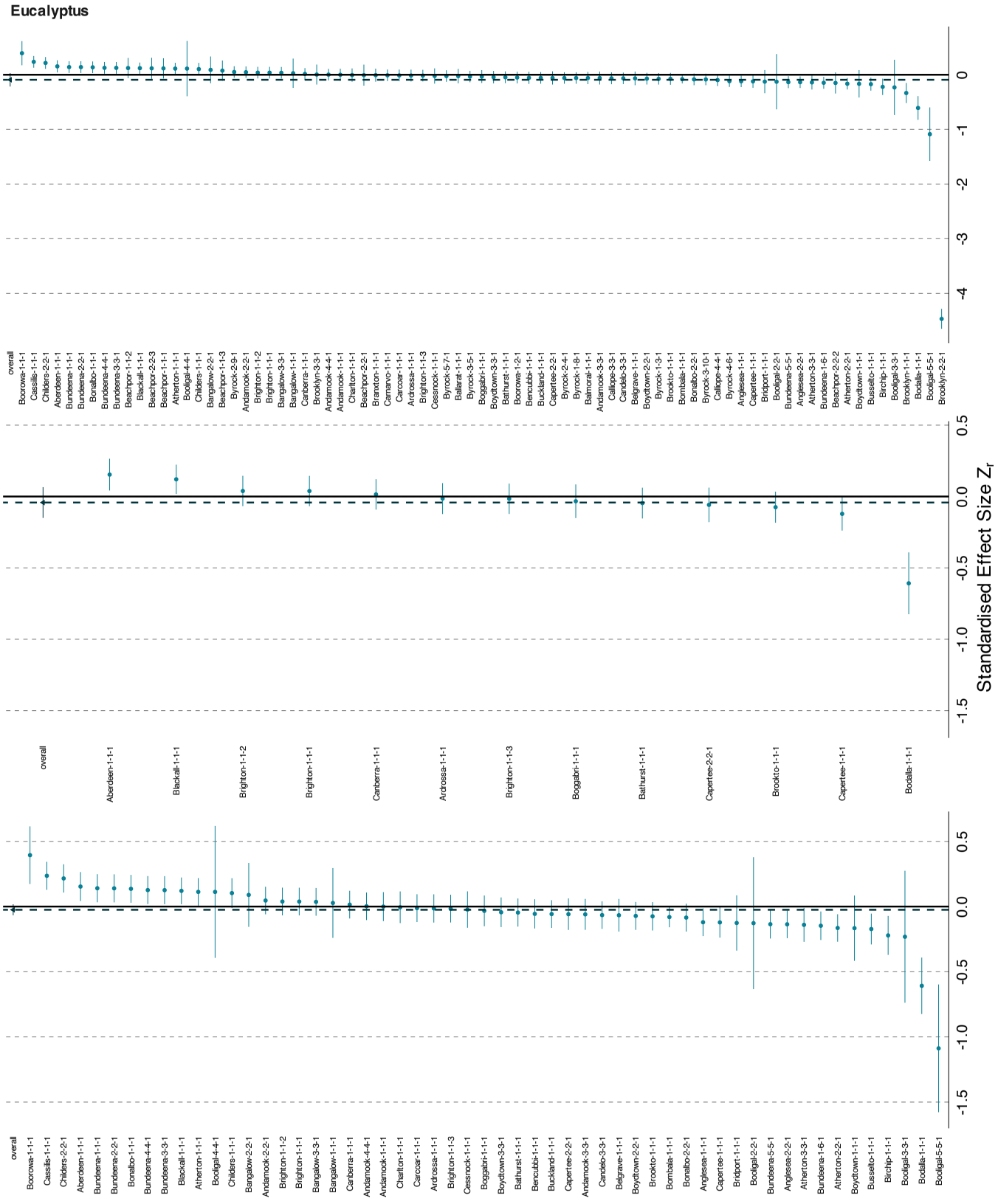


Figure A.6: *Forest plots of meta-analytic estimated standardised effect sizes (Z_r, blue circles) and their 95% confidence intervals for each effect size included in the meta-analysis model. The meta-analytic mean effect size is denoted by a black circle and a dashed vertical line, with error bars also representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0, indicating no relationship between the test variable and the response variable. The left side of each panel shows the analysis team names (anonymous arbitrary names assigned by us), each followed by three numbers. The first number is the submission ID (some analyst teams submitted results to us on >1 submission form), the second number is the analysis ID (some analyst teams included results of >1 analysis in a given submission), and the third number is the effect ID (some analysts submitted values for >1 effect per analysis). Thus, each row in each forest plot is uniquely identified, but it is possible to determine which effects come from which analyses and which analysis teams. The plots in the top row depict effects from analyses of blue tit data, and the bottom row plots depict effects from analyses of Eucalyptus data. The right-most plots depict all usable effect sizes. The plots on the left exclude effects from analysis sets that received at least one rating of "unpublishable" from peer reviewers, and the plots in the middle exclude effects from analysis sets that received at least one rating of either "unpublishable" or "major revision" from peer reviewers.*

Table A.4: *Estimated meta-analytic mean, standard error, and 95% confidence intervals, from analyses of the primary data set, the more conservative version of the dataset which excluded effects based on less reliable estimates of df, and both of these datasets with outliers removed.*

dataset	$\hat{\mu}$	SE[$\hat{\mu}$]	95%CI	statistic	p.value
Primary dataset					
blue tit	-0.35	0.03	[-0.41,-0.29]	-11.02	<0.001
Eucalyptus	-0.09	0.06	[-0.22,0.03]	-1.47	0.14
Conservative exclusions					
blue tit	-0.36	0.03	[-0.42,-0.29]	-10.77	<0.001
Eucalyptus	-0.11	0.07	[-0.24,0.03]	-1.55	0.12
Primary dataset, outliers removed					
blue tit	-0.36	0.03	[-0.42,-0.30]	-11.48	<0.001
Eucalyptus	-0.03	0.01	[-0.06,0.00]	-2.23	0.026
Conservative exclusions, outliers removed					
blue tit	-0.36	0.03	[-0.43,-0.30]	-11.38	<0.001
Eucalyptus	-0.04	0.02	[-0.07,-0.01]	-2.52	0.012

Post-hoc Analysis: Exploring the effect of excluding estimates in which we had reduced confidence For each dataset (blue tit, *Eucalyptus*), we created a second, more conservative version, that excluded effects based on estimates of *df* that we considered less reliable (Table A.4). We compared the outcomes of analyses of the primary dataset (constituted according to our registered plan) with the outcomes of analyses of the more conservative version of the dataset. We also compared results from analyses of both of these versions of the dataset to versions with our post-hoc removal of outliers described in the main text. Our more conservative exclusions (based on unreliable estimates of *df*) had minimal impact on the meta-analytic mean for both blue tit and *Eucalyptus* analyses, regardless of whether outliers were excluded (Table A.4).

Post-hoc Analysis: Exploring the effect of including only analyses conducted by analysis teams with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area The anonymous Team Identifiers in the reduced subset of "expert" or "highly proficient" analysts are exported internally in the *ManyEcoEvo* package as `expert_subset`. Analyses from the following teams are retained in the reduced subset: *Bell*, *Berr*, *Brim*, *Bruc*, *Burr*, *Byng*, *Cape*, *Clar*, *Clev*, *Alban*, *Alpha*, *Bargo*, *Berry*, *Bowen*, *Bulli*, *Aldgat*, *Alding*, *Anakie*, *Aramac*, *August*, *Bamaga*, *Barham*, *Barmah*, *Batlow*, *Beltan*, *Bethan*, *Beulah*, *Bindoo*, *Boonah*, *Bowral*, *Bright*, *Buchan*, *Burnie*,

Cairns, *Casino*, *Cattai*, *Adelong*, *Angasto*, *Antwerp*, *Arltung*, *Ashford*, *Babinda*, *Bargara*, *Barooga*, *Barraba*, *Belmont*, *Bemboka*, *Benalla*, *Bendigo*, *Berrima*, *Berwick*, *Beverle*, *Bicheno*, *Biloela*, *Birchip*, *Bombala*, *Bonalbo*, *Brookto*, *Bruthen*, *Buderim*, *Candelo*, *Capella*, *Carcoar*, *Carnama*, *Chewton*, *Anglesea*, *Ardrossa*, *Armidale*, *Atherton*, *Balaklav*, *Ballarat*, *Barellan*, *Belgrave*, *Berrigan*, *Binalong*, *Binnaway*, *Blackall*, *Boggabri*, *Bridport*, *Brooklyn*, *Buckland*, *Bundeena*, *Bungonia*, *Busselto*, *Calliope*, *Cardwell*, *Cassilis*, *Cessnock*, *Charlton*.

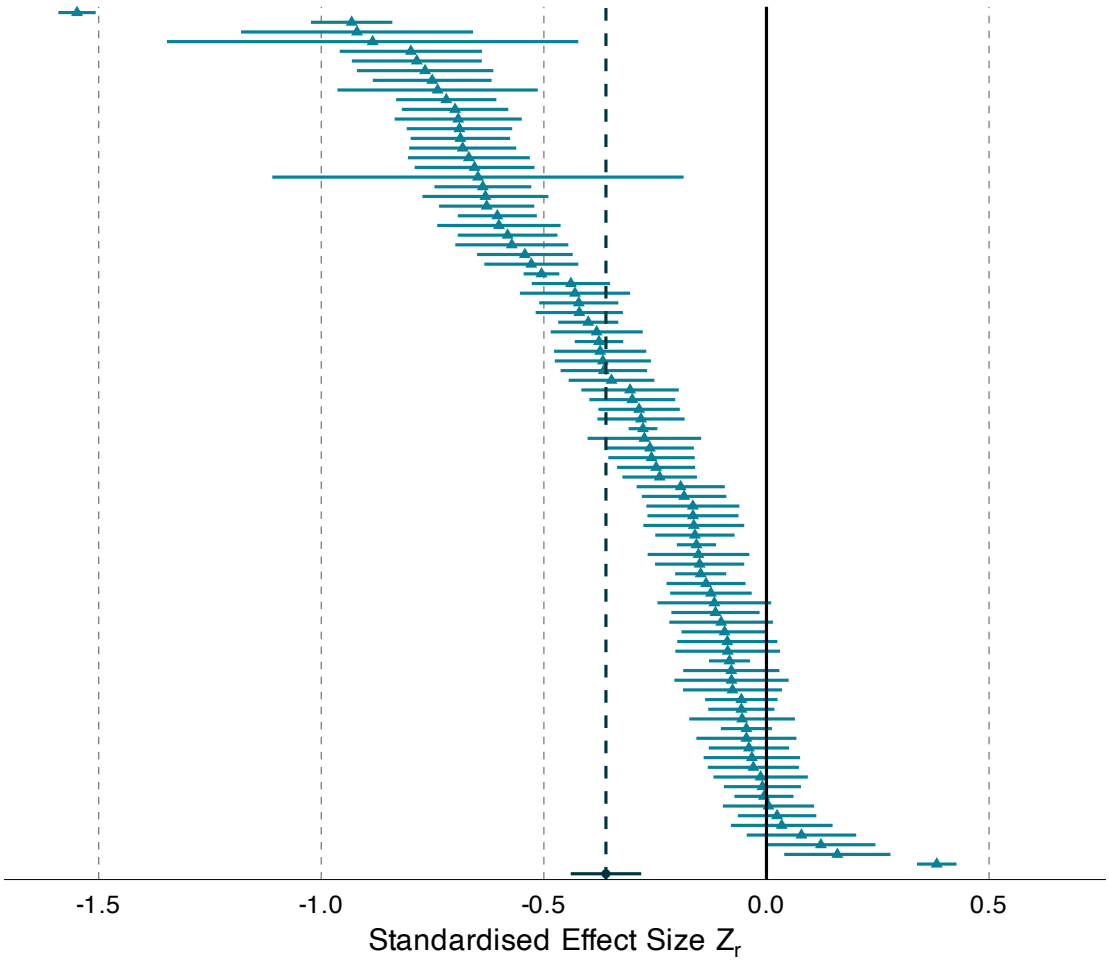
Post-hoc Analysis: Exploring the effect of excluding analyses of the blue tit dataset containing highly collinear predictor variables For the blue tit dataset, we created a subset of analyses that excluded effects based on analyses containing highly correlated predictor variables. Excluded analyses are exported internally in the *ManyEcoEvo* package as `ManyEcoEvo::collinearity_subset`. Analyses with the following identifiers are excluded in the reduced subset: *Armadal-1-1-1*, *Babinda-1-1-1*, *Babinda-2-2-1*, *Barham-1-1-1*, *Barham-2-2-1*, *Bega-1-1-1*, *Bega-1-1-2*, *Bega-2-2-1*, *Bega-2-2-2*, *Borde-1-1-1*, *Bruc-1-1-1*, *Caigun-1-1-1*, *Caigun-2-2-1*, *Adelong-1-1-1*, *Adelong-2-2-1*.

Out-of-sample Predictions y_i

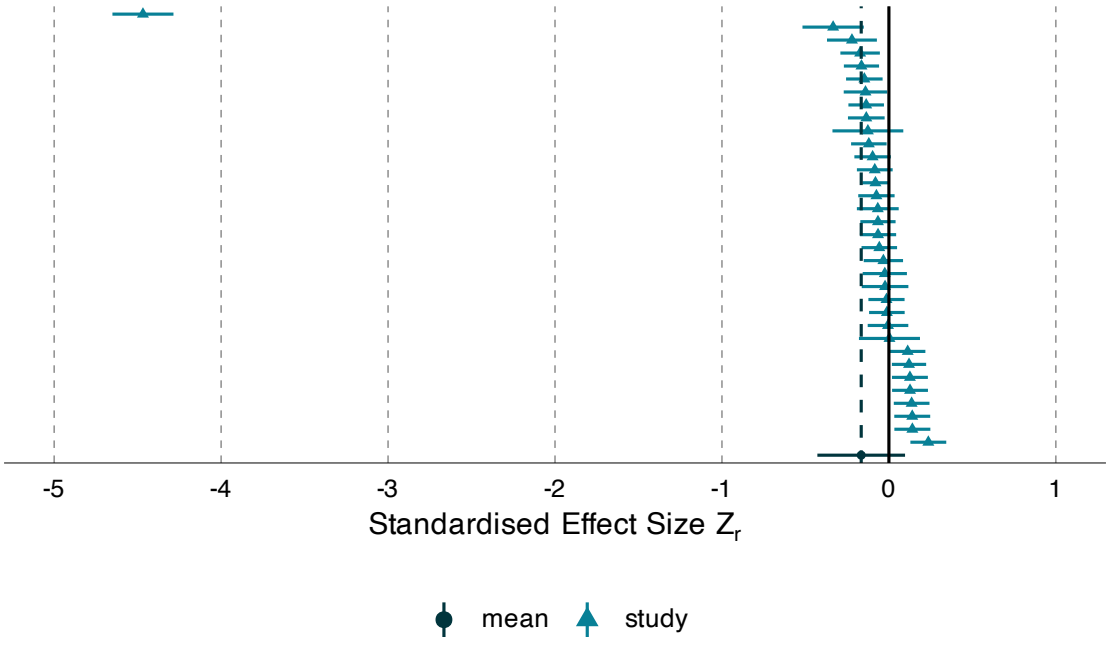
Excluded Analyses with Constructed Variables We standardised the y_i estimates and their standard errors for the blue tit analyses using the population mean and standard deviations of the corresponding dependent variable for that analysis, as shown in Equation A.1, using the function `ManyEcoEvo::Z_VZ_preds()`. Note that this is NOT the same process as standardizing the effect sizes Z_r. We used the mean and standard deviation of the relevant raw datasets as our 'population' parameters.

$$\begin{aligned} Z_j &= \frac{\mu_i - \bar{x}_j}{\mathrm{SD}_j} \\ \mathrm{VAR}_{Z_j} &= \frac{\mathrm{SE}_{\mu_i}}{\mathrm{SD}_j} \end{aligned} \tag{A.1}$$

Where μ is the population parameter taken from our original dataset for variable i, and $\bar{x}_j$ and SD_j are the out-of-sample point estimate values supplied for analysis j. SE_{μ_i} is the standard error of the population mean for variable i, while VAR_{Z_j} and Z_j are the standardised variance and mean estimate for analysis j. Note that for the response variables



(a) *Blue tit dataset analyses*



(b) Eucalyptus *dataset analyses*

Figure A.7: *Estimated meta-analytic effect size (Z_r, blue triangles), standard error, and 95% confidence intervals, from analyses of the primary data set with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area. Meta-analytic mean effect sizes are represented with black circles.*

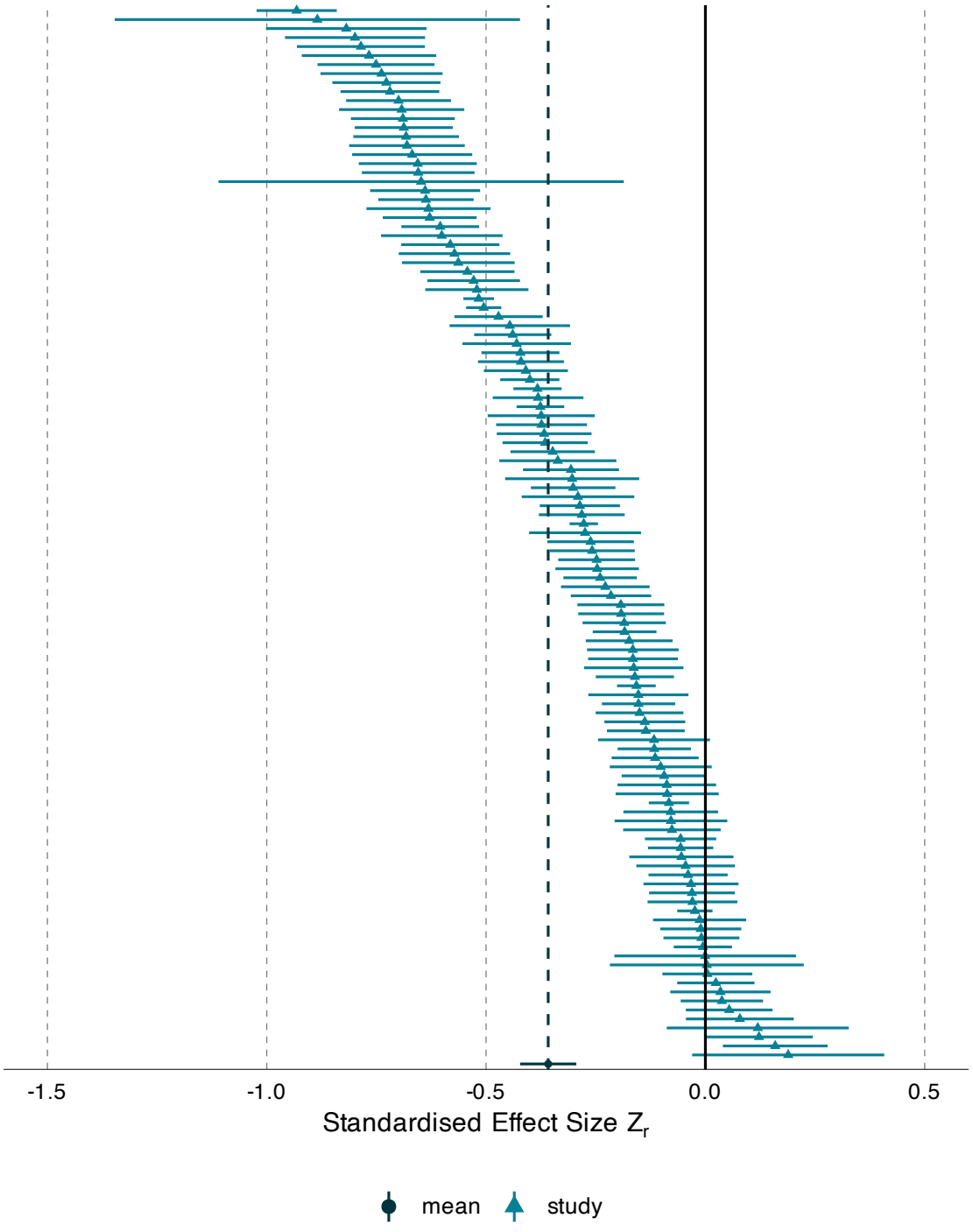


Figure A.8: *Forest plot of meta-analytic estimated effect-sizes (Z_r, blue triangles), standard error and 95% confidence intervals of blue tit analyses with highly collinear analyses removed. The meta-analytic mean for the reduced subset is denoted by the black circle, and a dashed vertical line, with error bars representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0.*

Table A.5: *Analyst-constructed variables and their inclusion in meta-analyses of out-of-sample predictions, y_i.*

Constructed Variable	Reconstructed for meta-analysis?
blue tit	
day_14_weight/day_14_tarsus_length	yes
day_14_weight/(day_14_tarsus_length^2)	yes
SMI	no
day_14_tarsus_length_group_deviation	no
day_14_weight_group_deviation	no
PC1.day_14_weight.day_14_tarsus_length	no
day_14_tarsus_length_deviation	no
residual_day14_weight	no
residual_day_14_weight_males	no
Eucalyptus	
euc_sdlgs0_2m	yes
euc_sdlgs_all	yes
euc_sdlgs>50cm	yes
small*0.25+medium*1.25+large*2.5	yes
average.proportion.of.plots.containing.at.least.one.euc.seedling.of.any.size	no

that were scaled-and-centered, or else mean-centred before model fitting, we do not need to standardise because these are already on the Z-scale. In doing so we make the assumption that analysts' data subsetting will have little effect on the outcomes. For some analyses of the blue tit dataset, analysts constructed their own unique response variables, which meant we needed to reconstruct these variables in order to calculate the population parameters. Unfortunately we were not able to re-construct all variables used by the analysts, as we were unable to reproduce the data required for their re-construction, e.g. we were unable to reproduce principal component analyses or fitted models for extracting residuals Table A.5. A total of 15 were excluded from out-of-sample meta-analysis, from 10 teams, including the following analysis identifiers: *Bruc-1-1-1*, *Clar-2-2-1*, *Clar-1-1-1*, *Batlow-1-1-1*, *Batlow-1-1-2*, *Bindoo-1-1-1*, *Bourke-1-1-1*, *Buchan-1-1-1*, *Arltung-4-4-1*, *Bargara-1-1-1*, *Bendigo-5-5-1*, *Booligal-3-3-1*, *Booligal-2-2-1*, *Booligal-4-4-1* and *Booligal-5-5-1*.

Non-truncated y_i Meta-analysis Forest Plot Below is the non-truncated version of Figure 2.4 showing a forest plot of the out-of-sample predictions, y_i, on the response-scale (stem counts), for *Eucalyptus* analyses, showing the full error bars of all model estimates.

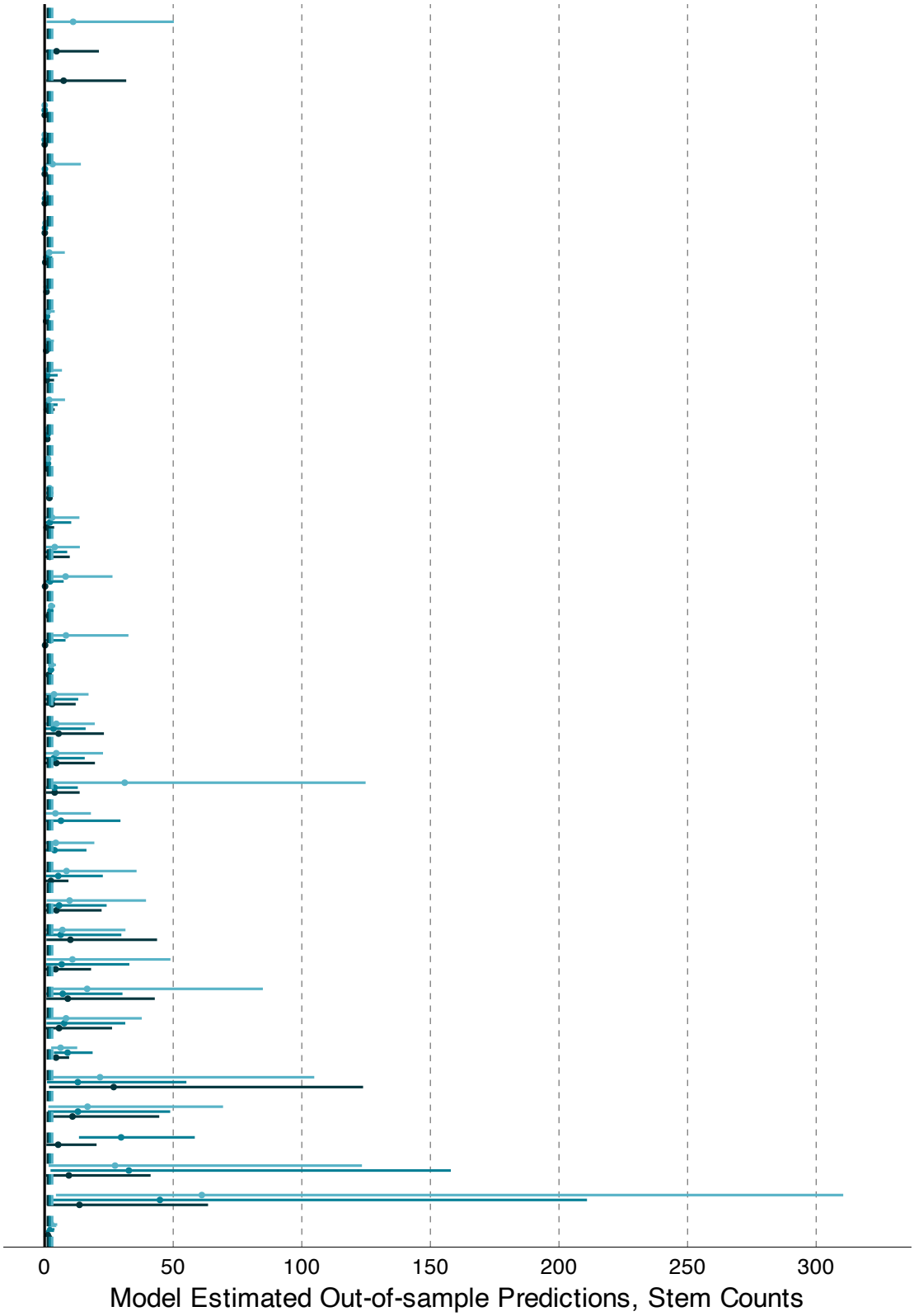


Figure A.9: *Forest plot of meta-analytic estimated out-of-sample predictions, y_i, on the response-scale (stem counts) for* Eucalyptus *analyses. Circles represent individual analysis estimates. Triangles represent the meta-analytic mean for each prediction scenario. Navy blue coloured points correspond to y_{25} scenario, blue coloured points correspond to the y_{50} scenario, while light blue points correspond to the y_{75} scenario. Error bars are 95% confidence intervals. Outliers (i.e. observations with mean estimates more than 3SD above the population parameter mean, see Appendix A.2.1) have been removed prior to model fitting.*

A.3 Explaining Variation in Deviation Scores

A.3.1 Box-Cox Transformation of Response Variable for Model Fitting

To aid in interpreting explanatory models where the response variable has been Box-Cox transformed, we plotted the transformation relationship for each of our analysis datasets (Figure A.10). Note that `timetk::step_box_cox()` directly optimises the estimation of the transformation parameter lambda, λ, using the "Guerrero" method such that λ minimises the coefficient of variation for sub-series of a numeric vector (Dancho & Vaughan, 2023, see `?timetk::step_box_cox()` for further details). Consequently, each dataset has its own unique value of λ, and therefore a unique transformation relationship.

A.3.2 Model Convergence and Singularity Problems

During model fitting, especially during fitting of models with random effects using `lme4::` (Bates et al., 2015), some models failed to converge while others were accompanied with console warnings of singular fit. However, the convergence checks from

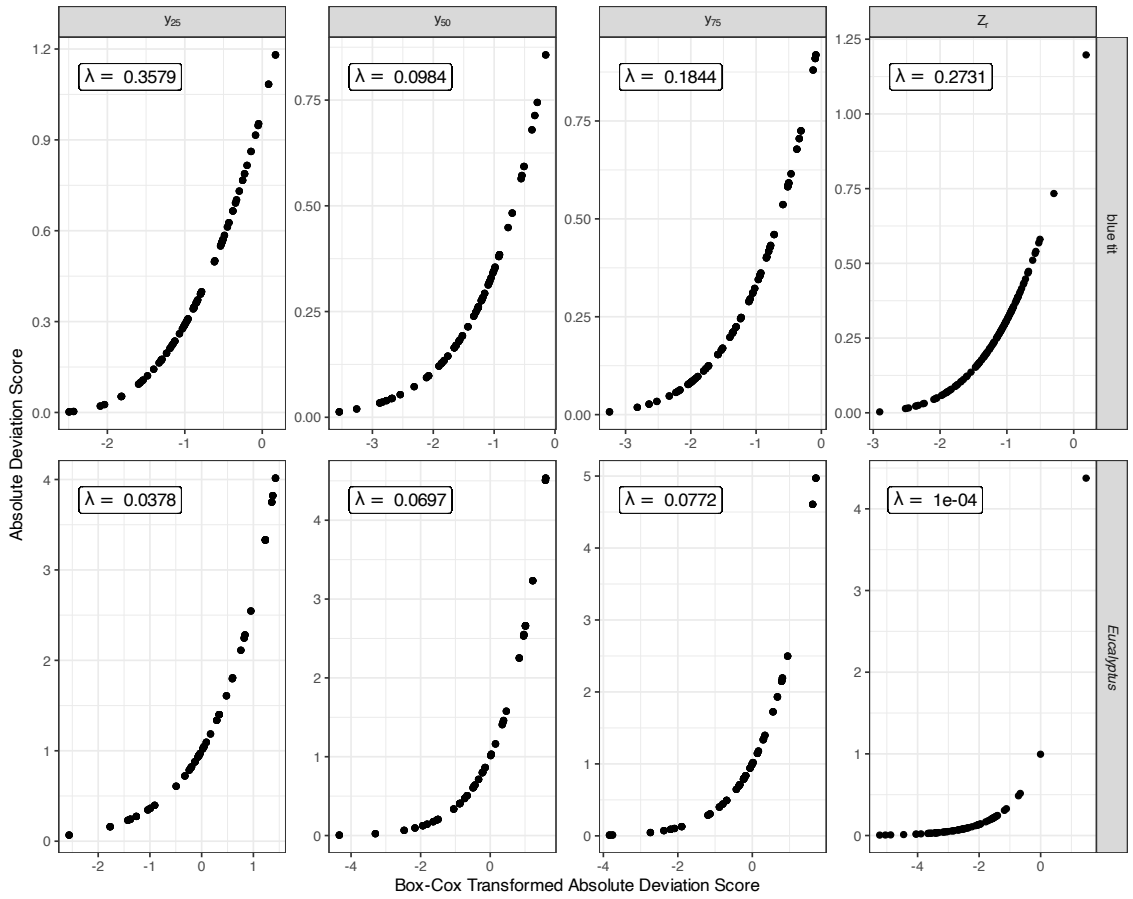


Figure A.10: *Box-Cox transformed absolute deviation scores plotted against (untransformed) absolute deviation scores.*

`lme4::` are known to be overly strict (see `?performance::check_convergence()` documentation for a discussion of this issue), consequently we checked for model warnings of convergence failure using the `performance::check_convergence()` function from the `performance::` package (Lüdecke, Ben-Shachar, et al., 2021). For all models we double-checked that they did not have singular fit by using `performance::check_singularity()`. Despite passing singularity checks with the `performance::` package, `parameters::parameters()` was unable to properly estimate SE and confidence intervals for the random effects of some models, which suggests singular fit. For all models we also checked whether the SE of random effects estimates could be calculated, and if not, marked these models as being singular. Analyses of singularity and convergence are presented throughout this document under the relevant section-heading for the analysis type and outcome, i.e. effect size (Z_r) or out-of-sample predictions (y_i).

A.3.3 Deviation Scores as Explained by Reviewer Ratings

Effect Sizes Z_r

Models pf deviation explained by categorical peer ratings all had singular fit or failed to converge for both blue tit and *Eucalyptus* datasets when random efects were included for both the effect ID and the reviewer ID (Table A.6). For the *Eucalyptus* dataset, when a random effect was included for effect ID only, the model failed to converge. The same was true for the blue tit dataset. As for the effect-size analysis, we included a random-effect for Reviewer ID only when fitting models of deviation explained by categorical peer ratings (See Table 2.6).

For models of deviation explained by continuous peer-review ratings, when including both random effects for effect ID and Reviewer ID model fits were singular for both datasets (Table A.6). The models passed the `performance::check_singularity()` check, however, however, the SD and CI could not be estimated by `parameters::model_parameters()` with a warning stating this was likely due to singular fit. For models with a random effect for effect ID, the same occurred for the blue tit dataset, whereas for the *Eucalyptus* dataset, the model did not converge at all. Consequently, for both blue tit and *Euclayptus* datasets, we fitted and analysed models of deviation explained by continuous peer review ratings with a random effect for Reviewer ID only (See Table 2.6).

Out-of-sample Predictions y_i

As for effect-size estimates Z_r, we encountered convergence and singularity problems when fitting models of deviation in out-of-sample predictions y_i explained by categorical peer ratings for both datasets (Table A.7). For all continuous models across both datasets, we encountered convergence and singularity problems when including random effects for both effect ID and Reviewer ID, as well as when including random effects for the effect ID only. In the latter case, for many prediction scenarios, across both blue tit and *Eucalyptus* datasets, estimated random effect coefficient CI's and SE could not be estimated. For models of deviation in out-of-sample predictions explained by continuous peer review ratings, when a random effect was included for effect ID only, CI's returned values of 0 for both bounds and model means estimated with `modelbased::estimate_means()` could not be reliably estimated and were equal for every peer-rating category (Table A.7). Consequently, we fitted models of deviation in out-of-sample predictions explained by continuous peer ratings with a random effect for Reviewer ID only (Table A.8). These model structures matched converging and non-singular model structures for effect-size estimates Z_r (Table A.6).

Table A.6: *Singularity and convergence checking outcomes for models of deviation in effect-sizes Z_r explained by peer-review ratings for different random effect structures. Problematic checking outcomes are highlighted in red.*

Fixed Effect	Random Effects	Model converged?	Singular Fit?	Can random effect SE be calculated?	Can random effect 95% CI be calculated?
blue tit					
publishable_as_is	Reviewer ID	yes	no	yes	yes
	Study ID, Reviewer ID	yes	no	yes	**no**
	Study ID	**no**	—	—	—
rate_analysis	Reviewer ID	yes	no	yes	yes
	Study ID, Reviewer ID	yes	no	**no**	**no**
	Study ID	yes	no	**no**	**no**
Eucalyptus					
publishable_as_is	Reviewer ID	yes	no	yes	yes
	Study ID, Reviewer ID	yes	**yes**	**no**	**no**
	Study ID	**no**	—	—	—
rate_analysis	Reviewer ID	yes	no	yes	yes
	Study ID, Reviewer ID	yes	no	**no**	**no**
	Study ID	**no**	—	—	—

Table A.7: *Singularity and convergence checking outcomes for models of deviation in out-of-sample predictions y_i explained by peer-review ratings for different random effect structures. Problematic checking outcomes are highlighted in red.*

	Random Effects	Prediction Scenario	Model converged?	Singular Fit?	Can random effects SE be calculated?	Can random effect 95% CI be calculated?
Categorical Peer Rating						
Eucalyptus	Reviewer ID	y_{25}	yes	no	yes	yes
	Reviewer ID	y_{50}	yes	no	yes	yes
	Reviewer ID	y_{75}	yes	**yes**	**no**	yes
Eucalyptus	Study ID, Reviewer ID	y_{25}	yes	no	yes	yes
	Study ID, Reviewer ID	y_{50}	yes	no	yes	yes

	Study ID, Reviewer ID	y_{75}	yes	**yes**	**no**	yes
Eucalyptus	Study ID	y_{25}	yes	no	yes	yes
	Study ID	y_{50}	yes	no	yes	yes
	Study ID	y_{75}	yes	no	yes	yes
blue tit	Reviewer ID	y_{25}	yes	**yes**	**no**	yes
	Reviewer ID	y_{50}	yes	no	yes	yes
	Reviewer ID	y_{75}	yes	no	yes	yes
blue tit	Study ID, Reviewer ID	y_{25}	yes	no	yes	yes
	Study ID, Reviewer ID	y_{50}	yes	**yes**	**no**	yes
	Study ID, Reviewer ID	y_{75}	yes	no	yes	yes
blue tit	Study ID	y_{25}	yes	no	yes	yes
	Study ID	y_{50}	yes	no	yes	yes
	Study ID	y_{75}	yes	no	yes	yes
Continuous Peer Rating						
Eucalyptus	Reviewer ID	y_{25}	yes	no	yes	yes
	Reviewer ID	y_{50}	yes	**yes**	**no**	yes
	Reviewer ID	y_{75}	yes	**yes**	**no**	yes
Eucalyptus	Study ID, Reviewer ID	y_{25}	yes	no	**no**	yes
	Study ID, Reviewer ID	y_{50}	**no**	—	—	—
	Study ID, Reviewer ID	y_{75}	yes	no	**no**	yes
Eucalyptus	Study ID	y_{25}	yes	no	**no**	yes
	Study ID	y_{50}	yes	no	**no**	yes
	Study ID	y_{75}	yes	no	yes	yes
blue tit	Reviewer ID	y_{25}	yes	**yes**	**no**	yes
	Reviewer ID	y_{50}	yes	no	yes	yes
	Reviewer ID	y_{75}	yes	no	yes	yes
blue tit	Study ID, Reviewer ID	y_{25}	yes	no	**no**	yes
	Study ID, Reviewer ID	y_{50}	yes	no	**no**	yes
	Study ID, Reviewer ID	y_{75}	yes	no	**no**	yes
blue tit	Study ID	y_{25}	yes	no	**no**	yes
	Study ID	y_{50}	yes	no	**no**	yes
	Study ID	y_{75}	yes	no	yes	yes

Table A.8: *Singularity and convergence checking for models of deviation in out-of-sample-predictions y_i explained by peer-ratings.*

	Prediction Scenario	Singular Fit?	Model converged?	Can random effects SE be calculated?	Can random effect 95% CI be calculated?
Deviation explained by continuous ratings					
blue tit	y_{25}	**yes**	yes	**no**	**no**
	y_{50}	no	yes	yes	yes
	y_{75}	no	yes	yes	yes
Eucalyptus	y_{25}	no	yes	yes	yes
	y_{50}	**yes**	yes	**no**	**no**
	y_{75}	**yes**	yes	**no**	**no**
Deviation explained by categorical ratings					
blue tit	y_{25}	**yes**	yes	**no**	**no**
	y_{50}	no	yes	yes	yes
	y_{75}	no	yes	yes	yes
Eucalyptus	y_{25}	no	yes	yes	yes
	y_{50}	no	yes	yes	yes
	y_{75}	**yes**	yes	**no**	**no**

We fitted the same deviation models on the out-of-sample-predictions dataset that we fitted for the effect-size dataset. However, while all models of deviation explained by categorical peer-ratings converged, the following datasets and prediction scenarios suffered from singular fit: blue tit - y_{25}, *Eucalyptus* - y_{75} (Table A.8). Models of deviation explained by *continuous* ratings all converged, however models for the out-of-sample predictions model fit was singular. Similarly to the effect-size (Z_r) dataset, SD and CI could not be estimated for random effects in some models (Table A.8), consequently we interpreted this to mean the models had singular fit (See Appendix A.3.3). Results of all deviation models are therefore presented only for models with non-singular fit, and that converged (Table A.8).

Group means and 95% confidence intervals for different categories of peer-review rating are all overlapping (Figure A.11). The fixed effect of peer review rating also explains virtually no variability in deviation scores for out-of-sample predictions y_i (Table A.8).

There was a lack of any clear relationships between quantitative review scores and y_i deviation scores (Table A.16). Plots of these relationships indicated either no relationship or extremely weak positive relationships (Figure A.12). Recall that positive relationships mean that as review scores became more favorable, the deviation from the meta-analytic mean increased, which is surprising. Because almost no variability in y_i deviation score was

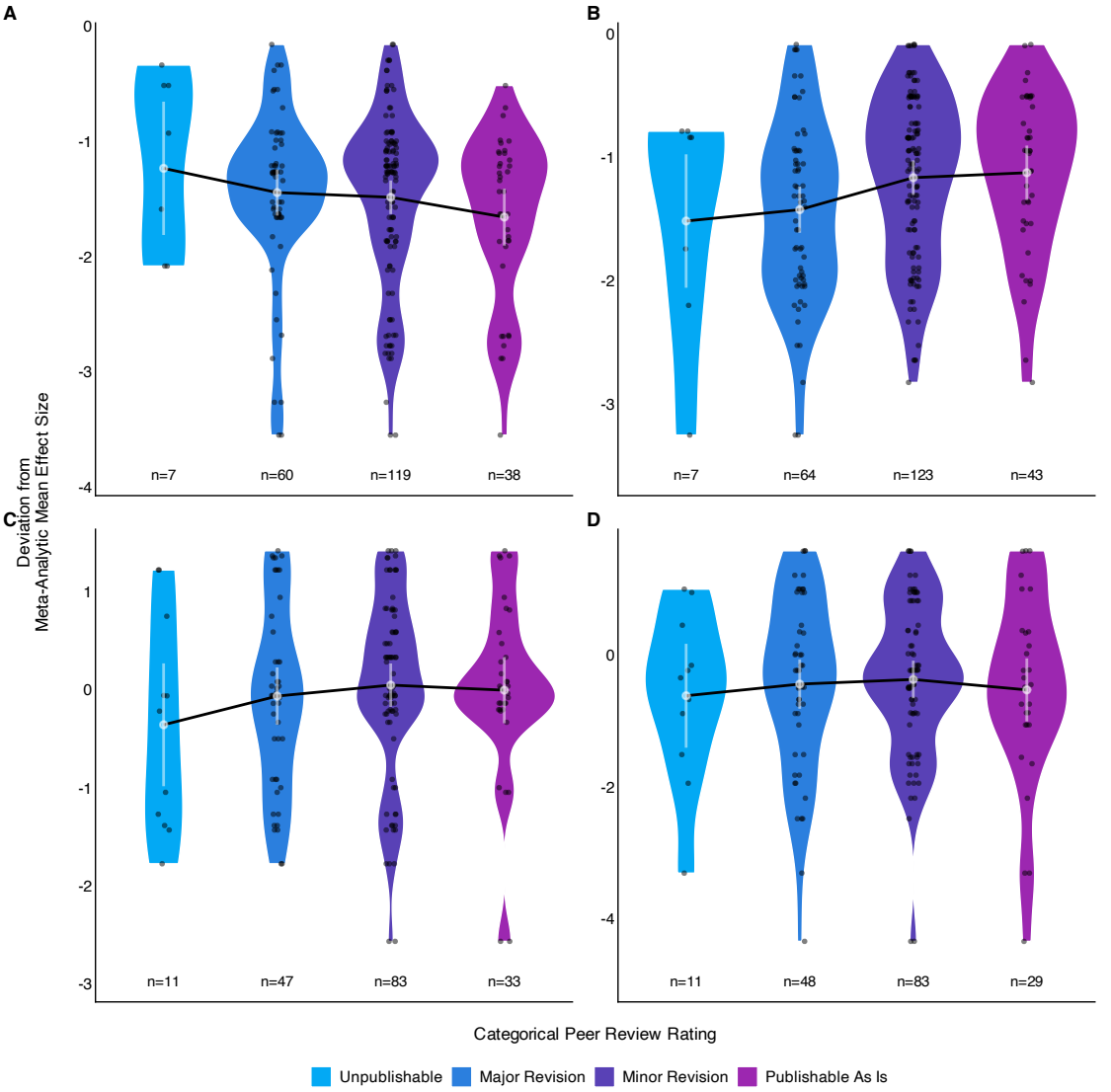


Figure A.11: *Violin plot of Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of categorical peer-review ratings ratings. Note that higher (less negative) values of the deviation score result from greater deviation from the meta-analytic mean. Grey points for each rating group denote model-estimated marginal mean deviation, and error bars denote 95% CI of the estimate.* ***A****: Blue tit, y50,* ***B****: Blue tit, y75,* ***C****:* Eucalyptus*, y25,* ***D****:* Eucalyptus*, y50.*

explained by reviewer ratings (Table A.16), this pattern does not appear to merit further consideration.

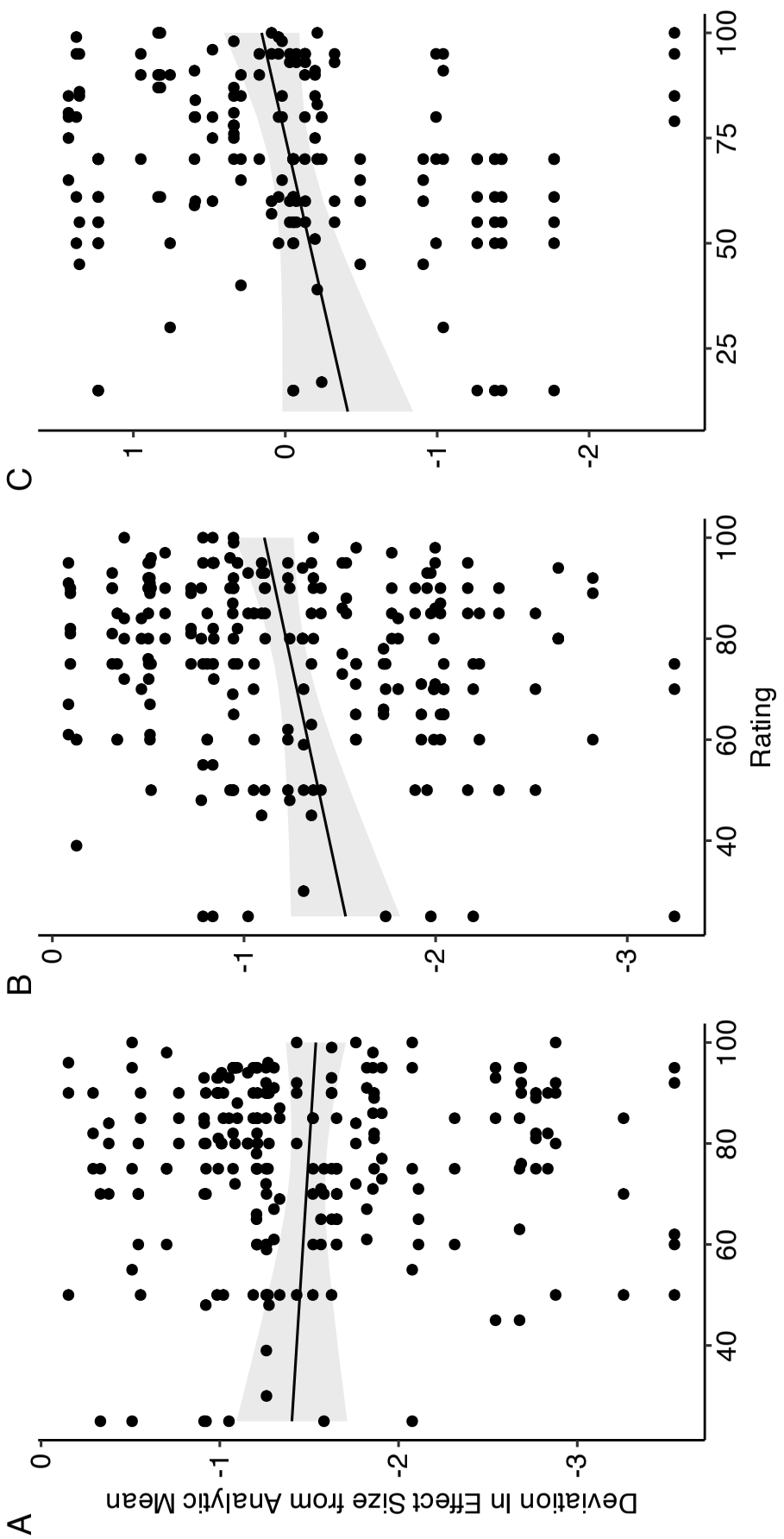


Figure A.12: *Scatterplots explaining Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of continuous peer-review ratings. Note that higher (less negative) values of the deviation score result from greater deviation from the meta-analytic mean.* ***A****: Blue tit, y50,* ***B****: Blue tit, y75,* ***C****:* Eucalyptus, *y25.*

Table A.9: *Parameter estimates for univariate models of Box-Cox transformed deviation from the mean y_i estimate as a function of categorical peer-review rating, continuous peer-review rating, and Sorensen's index for blue tit and* Eucalyptus *analyses, and also for the inclusion of random effects for* Eucalyptus *analyses.*

	Prediction Scenario	Parameter	Random Effect	Coefficient	SE	95%CI	t	df	p
Deviation explained by categorical ratings									
Eucalyptus	y_{25}	(Intercept)		-0.35	0.32	[-0.98,0.28]	-1.1	168	0.3
	y_{25}	Publishable with major revision		0.29	0.35	[-0.39,0.98]	0.84	168	0.4
	y_{25}	Publishable with minor revision		0.41	0.34	[-0.26,1.07]	1.21	168	0.2
	y_{25}	Publishable as is		0.35	0.36	[-0.35,1.06]	0.99	168	0.3
	y_{25}	SD (Intercept)	Reviewer ID	0.23	0.15	[0.07,0.79]			
	y_{25}	SD (Observations)	Residual	0.92	0.06	[0.82,1.04]			
Eucalyptus	y_{50}	(Intercept)		-0.61	0.4	[-1.4,0.18]	-1.52	165	0.13
	y_{50}	Publishable with major revision		0.18	0.44	[-0.69,1.05]	0.4	165	0.7
	y_{50}	Publishable with minor revision		0.25	0.42	[-0.59,1.09]	0.58	165	0.6
	y_{50}	Publishable as is		0.09	0.47	[-0.83,1.01]	0.19	165	0.8
	y_{50}	SD (Intercept)	Reviewer ID	0.12	0.42	$[1.50 \times 10^{-4}, 1.02 \times 10^{2}]$			
	y_{50}	SD (Observations)	Residual	1.29	0.08	[1.14,1.46]			
blue tit	y_{50}	(Intercept)		-1.23	0.29	[-1.81,-0.65]	-4.18	218	<0.001
	y_{50}	Publishable with major revision		-0.21	0.3	[-0.81,0.39]	-0.69	218	0.5
	y_{50}	Publishable with minor revision		-0.25	0.3	[-0.85,0.34]	-0.83	218	0.4
	y_{50}	Publishable as is		-0.42	0.32	[-1.05,0.21]	-1.33	218	0.2
	y_{50}	SD (Intercept)	Reviewer ID	0.22	0.09	[0.11,0.47]			
	y_{50}	SD (Observations)	Residual	0.71	0.04	[0.64,0.8]			
blue tit	y_{75}	(Intercept)		-1.51	0.27	[-2.05,-0.97]	-5.52	231	<0.001
	y_{75}	Publishable with major revision		0.09	0.28	[-0.46,0.64]	0.33	231	0.7
	y_{75}	Publishable with minor revision		0.35	0.28	[-0.2,0.9]	1.25	231	0.2
	y_{75}	Publishable as is		0.39	0.29	[-0.19,0.97]	1.33	231	0.2
	y_{75}	SD (Intercept)	Reviewer ID	0.29	0.06	[0.18,0.44]			
	y_{75}	SD (Observations)	Residual	0.63	0.03	[0.57,0.7]			
Deviation explained by continuous ratings									
Eucalyptus	y_{25}	(Intercept)		-0.46	0.26	[-0.97,0.06]	-1.76	170	0.081

	y_{25}	RateAnalysis		6.14×10^{-3}	3.54×10^{-3}	$[-8.61 \times 10^{-4}, 1.31 \times 10^{-2}]$	1.73	170	0.085
	y_{25}	SD (Intercept)	Reviewer ID	0.12	0.21	$[4.27 \times 10^{-3}, 3.64]$			
	y_{25}	SD (Observations)	Residual	0.93	0.06	[0.82,1.05]			
blue tit	y_{50}	(Intercept)		-1.35	0.24	[-1.82,-0.88]	-5.65	220	<0.001
	y_{50}	RateAnalysis		-1.82×10^{-3}	3.05×10^{-3}	$[-7.83 \times 10^{-3}, 4.18 \times 10^{-3}]$	-0.6	220	0.6
	y_{50}	SD (Intercept)	Reviewer ID	0.24	0.08	[0.12,0.47]			
	y_{50}	SD (Observations)	Residual	0.71	0.04	[0.64,0.79]			
blue tit	y_{75}	(Intercept)		-1.66	0.22	[-2.1,-1.23]	-7.52	233	<0.001
	y_{75}	RateAnalysis		5.62×10^{-3}	2.79×10^{-3}	$[1.22 \times 10^{-4}, 1.11 \times 10^{-2}]$	2.01	233	0.045
	y_{75}	SD (Intercept)	Reviewer ID	0.3	0.06	[0.2,0.45]			
	y_{75}	SD (Observations)	Residual	0.63	0.03	[0.57,0.7]			
Deviation explained by Sorensen's index									
Eucalyptus	y_{25}	(Intercept)		-0.3	1.48	[-3.2,2.59]	-0.21	36	0.8
	y_{25}	Mean Sorensen's index		0.44	2.19	[-3.86,4.74]	0.2	36	0.8
Eucalyptus	y_{50}	(Intercept)		-1.32	2.1	[-5.43,2.79]	-0.63	36	0.5
	y_{50}	Mean Sorensen's index		1.32	3.16	[-4.87,7.51]	0.42	36	0.7
Eucalyptus	y_{75}	(Intercept)		-0.71	1.78	[-4.19,2.78]	-0.4	36	0.7
	y_{75}	Mean Sorensen's index		0.34	2.71	[-4.96,5.64]	0.12	36	>0.9
blue tit	y_{25}	(Intercept)		-0.77	0.6	[-1.94,0.4]	-1.29	61	0.2
	y_{25}	Mean Sorensen's index		-0.23	1.04	[-2.27,1.82]	-0.22	61	0.8
blue tit	y_{50}	(Intercept)		-0.43	0.73	[-1.86,0.99]	-0.6	58	0.6
	y_{50}	Mean Sorensen's index		-1.77	1.27	[-4.26,0.71]	-1.4	58	0.2
blue tit	y_{75}	(Intercept)		-1.72	0.74	[-3.16,-0.28]	-2.34	61	0.019
	y_{75}	Mean Sorensen's index		0.78	1.28	[-1.73,3.29]	0.61	61	0.5
Deviation explained by inclusion of random effects									
Eucalyptus	y_{25}	(Intercept)		-0.53	0.26	[-1.05,-0.02]	-2.03	36	0.042
	y_{25}	Mixed model		0.74	0.31	[0.13,1.35]	2.37	36	0.018
Eucalyptus	y_{50}	(Intercept)		-0.57	0.4	[-1.36,0.21]	-1.43	36	0.2
	y_{50}	Mixed model		0.18	0.48	[-0.75,1.11]	0.38	36	0.7
Eucalyptus	y_{75}	(Intercept)		-0.35	0.39	[-1.11,0.41]	-0.9	36	0.4
	y_{75}	Mixed model		-0.19	0.46	[-1.1,0.71]	-0.41	36	0.7

A.3.4 Deviation Scores as Explained by the Distinctiveness of Variables in Each Analysis

Out-of-sample Predictions y_i

Given the convergence and singularity issues encountered with most other analyses, we also checked for convergence and singularity issues in models of deviation explained by Sorensen’s similarity index for y_i estimates (Table A.10). All models fitted without problem.

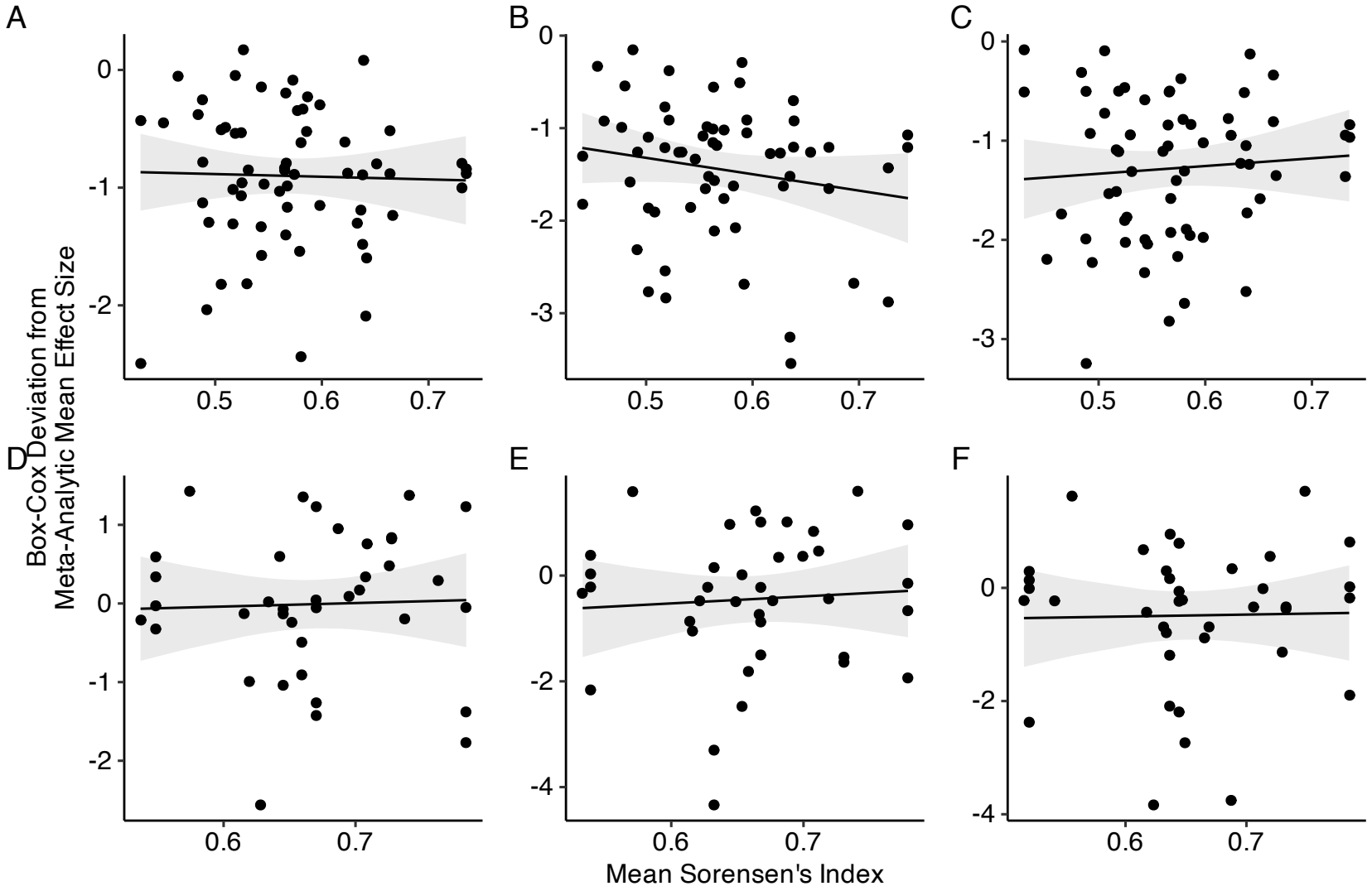


Figure A.13: *Scatter plots examining Box-Cox transformed deviation from the meta-analytic mean for y_i estimates as a function of Sorensen's similarity index. Note that higher (less negative) values of the deviation score result from greater deviation from the meta-analytic mean (Figure A.10).* ***A****: blue tit, y25,* ***B****: blue tit, y50,* ***C****: blue tit, y75,* ***D****: Eucalyptus, y25,* ***E****: Eucalyptus, y50,* ***F****: Eucalyptus, y75.*

Table A.10: *Singularity and convergence checks for models of deviation explained by Sorensen's similarity index and inclusion of random effects for out-of-sample predictions, y_i. Models of Deviation explained by inclusion of random effects are not presented for blue tit analyses because the number of models not using random effects was less than our preregistered threshold.*

	Prediction Scenario	Singular Fit?	Model converged?	Can random effect SE be calculated?	Can random effect CI be calculated?
Deviation explained by Sorensen's index					
blue tit	y_{25}	no	yes	yes	yes
	y_{50}	no	yes	yes	yes
	y_{75}	no	yes	yes	yes
Eucalyptus	y_{25}	no	yes	yes	yes
	y_{50}	no	yes	yes	yes
	y_{75}	no	yes	yes	yes
Deviation explained by inclusion of random effects					
Eucalyptus	y_{25}	no	yes	yes	yes
	y_{50}	no	yes	yes	yes
	y_{75}	no	yes	yes	yes

We checked the fitted models for the inclusion of random effects for the *Eucalyptus* dataset, and for models of deviation explained by Sorensen's similarity index for y_i estimates (Table A.10). All models converged, and no singular fits were encountered.

A.3.5 Deviation Scores as Explained by the Inclusion of Random Effects

Out-of-sample Predictions y_i

Only 1 of the blue tit out-of-sample analyses y_i included random effects, which was below our preregistered threshold of 5 for running the models of Box-Cox transformed deviation from the meta-analytic mean explained by the inclusion of random-effects. However, 14 *Eucalyptus* analyses included in the out-of-sample y_i results included only fixed effects, which crossed our pre-registered threshold.

Consequently, we performed this analysis for the *Eucalyptus* dataset only, here we present results for the out-of-sample prediction y_i results. There is inconsistent evidence of somewhat higher Box-Cox-transformed deviation values for models including a random effect, meaning the analyses of the *Eucalyptus* dataset that included random effects averaged slightly higher deviation from the meta-analytic mean out-of-sample estimate in the relevant prediction scenario. This is most evident for the y_{25} predictions which both shows the greatest difference in Box-Cox transformed deviation values (Figure A.14) and explains the most variation in y_i deviation score (Table A.9).

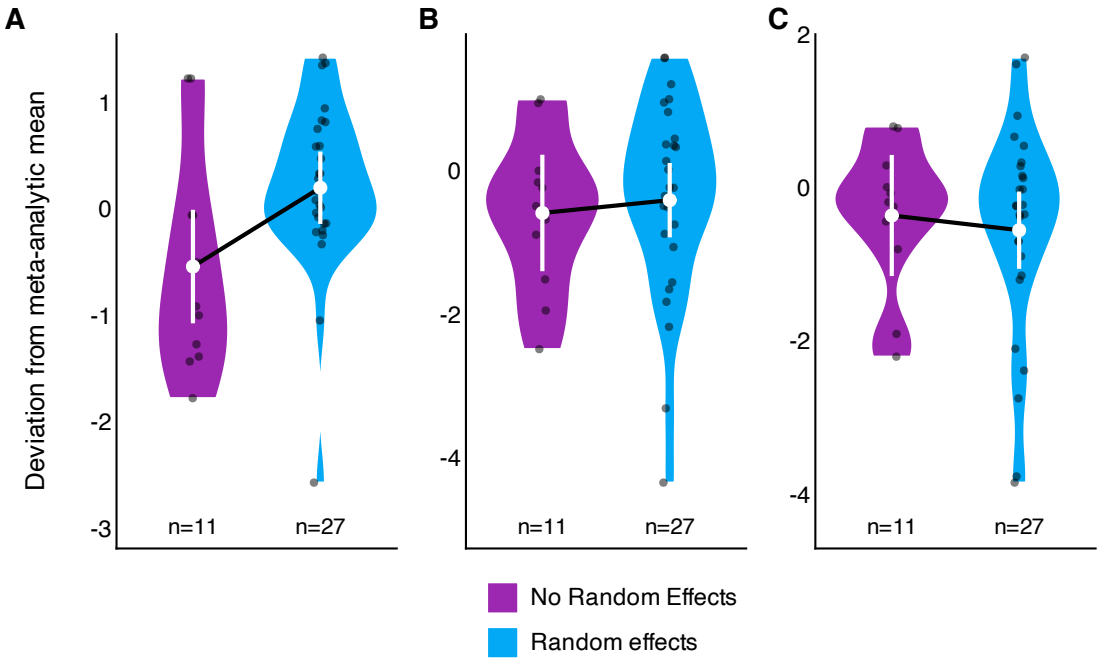


Figure A.14: *Violin plot of Box-Cox transformed deviation from meta-analytic mean as a function of presence or absence of random effects in the analyst's model. White points for each rating group denote model-estimated marginal mean deviation, and error bars denote 95% CI of the estimate. Note that higher (less negative) values of Box-Cox transformed deviation result from greater deviation from the meta-analytic mean.* ***A:*** Eucalyptus, *y25,* ***B:*** Eucalyptus, *y50,* ***C:*** Eucalyptus, *y75.*

A.3.6 Multivariate Analysis

Effect Sizes Z_r

The multivariate models did a poor job of explaining how different from the meta-analytic mean each analysis would be. For the blue tit analyses the R^2 value for the whole model was 0.11 and for the fixed effects component was 0.01, and the residual standard deviation for the model was 0.5. Further, all of the fixed effects had 95% confidence intervals that overlaped 0. This evidence is all consistent with none of the predictor variables in this model (continuous review rating, categorical review rating, distinctiveness of variables included) having any meaningful effect on how far Z_r estimates fell from the meta-analytic mean for the blue tit analyses. The pattern is largely similar for the *Eucalyptus* multivariate analysis, in which R^2 for the whole model was 0.13 and for the fixed effects component was 0.02, and the residual standard deviation for the model was 1.06. There is somewhat more of a hint of a pattern when examining the parameter estimates from the *Eucalyptus* analysis. In the case of the fixed effect of categorical reviewer ratings, analyses that were reviewed as 'publishable as is' and 'publishable with major revisions' appeared to produce results

Table A.11: *Parameter estimates from models explaining Box-Cox transformed deviation scores from the mean Z_r as a function of continuous and categorical peer-review ratings in multivariate analyses. Standard Errors (SE), 95% confidence intervals (95% CI) are reported for all estimates, while t values, degrees of freedom (df) and p-values are presented for fixed-effects.*

Parameter	Random Effect	Coefficient	SE	95% CI	t	df	*p*
blue tit							
(Intercept)		-1.24	0.24	[-1.71,-0.77]	-5.16	465	p<0.001
RateAnalysis		-2.84×10^{-3}	0	$[-7.81 \times 10^{-3}, 2.13 \times 10^{-3}]$	-1.12	465	p=0.3
Publishable as is		0.06	0.2	[-0.33,0.45]	0.3	465	p=0.8
Publishable with major revision		-0.09	0.14	[-0.36,0.18]	-0.67	465	p=0.5
Publishable with minor revision		-0.02	0.17	[-0.36,0.31]	-0.14	465	p=0.9
Mean Sorensen's index		0.33	0.27	[-0.2,0.87]	1.22	465	p=0.2
SD (Intercept)	Reviewer ID	0.16	0.03	[0.11,0.25]			
SD (Observations)	Residual	0.5	0.02	[0.46,0.53]			
Eucalyptus							
(Intercept)		-2.8	0.77	[-4.32,-1.27]	-3.61	337	p<0.001
RateAnalysis		-0.01	0.01	$[-2.27 \times 10^{-2}, 1.74 \times 10^{-3}]$	-1.69	337	p=0.093
Publishable as is		0.76	0.54	[-0.3,1.82]	1.42	337	p=0.2
Publishable with major revision		0.65	0.36	[-0.06,1.35]	1.79	337	p=0.074
Publishable with minor revision		0.55	0.44	[-0.32,1.41]	1.24	337	p=0.2
Mean Sorensen's index		0.36	0.91	[-1.43,2.15]	0.39	337	p=0.7
Random included		0.19	0.2	[-0.2,0.59]	0.97	337	p=0.3
SD (Intercept)	Reviewer ID	0.38	0.09	[0.24,0.61]			
SD (Observations)	Residual	1.06	0.04	[0.98,1.15]			

Table A.12: *Model summary metrics for multivariate meta-analysis models. σ is the residual standard deviation, ICC is the intra-class correlation coefficient, and R^2_M and R^2_C are the marginal and conditional R^2, respectively.*

Dataset	$R^2_{\text{Conditional}}$	R^2_{Marginal}	ICC	RMSE	σ
blue tit	0.11	0.01	0.1	0.48	0.5
Eucalyptus	0.13	0.02	0.11	1.02	1.06

more different from the meta-analytic mean than those that were in the reference class of 'deeply flawed and unpublishable'. However, the estimates are very uncertain (*Eucalyptus* fixed effect for 'publishable as is' 0.76 (95% CI -0.29,1.81), and for 'publishable with major revision' 0.06 (95% CI -0.33,0.45)). Further, the collinearity between the categorical and continuous ratings make interpretation of effects involving either of these two variables unclear, and so we recommend against interpreting the pattern observed here. We report this analysis only for the sake of transparency.

Table A.13: *Singularity and convergence checks for all combinations of random effects specifications trialled for across subsets of out-of-sample predictions y_i from multivariate models.*

	Random Effects	Model converged?	Singular Fit?	Can random effects SE be calculated?	Can random effect CI be calculated?
blue tit					
y_{25}	Reviewer ID	yes	yes	no	no
y_{50}	Reviewer ID	yes	no	yes	yes
y_{75}	Reviewer ID	yes	no	yes	yes
y_{25}	Study ID, Reviewer ID	yes	yes	no	no
y_{50}	Study ID, Reviewer ID	yes	no	no	no
y_{75}	Study ID, Reviewer ID	yes	yes	no	no
y_{25}	Study ID	yes	no	no	no
y_{50}	Study ID	yes	no	no	no
Eucalyptus					
y_{25}	Reviewer ID	yes	yes	no	no
y_{50}	Reviewer ID	yes	no	yes	yes
y_{75}	Reviewer ID	yes	no	yes	yes
y_{25}	Study ID, Reviewer ID	yes	no	no	no
y_{50}	Study ID, Reviewer ID	yes	yes	no	no
y_{25}	Study ID	yes	no	no	no
y_{50}	Study ID	yes	no	no	no

Out-of-sample Predictions y_i

For the blue tit analyses, the only models that did converge, which were not singular and that had estimable random effect variances were the y_{50} and y_{75} prediction scenarios with Reviewer ID as the model random effect (Table A.13). Of the different random effects structures we trialled for the *Eucalyptus* analyses, only the model that included Reviewer ID as the random effect successfully fitted to the y_{50} and y_{75} prediction scenarios, with other models either failing to converge due to complete separation (`lme4::` error: `Downdated VtV is not positive definite`, see github.com/lme4/lme4/issues/483).

Consequently, we deviated from our intended plan of using random effects for both Effect ID and Reviewer ID, instead using a single random effect for Reviewer ID for the y_{50} and y_{75} prediction scenarios for both blue tit and *Eucalyptus* datasets (Table A.15, Table A.14).

A.3.7 Model Summary Metrics for Out-of-sample Predictions y_i

See Table A.16.

Table A.14: *Parameter estimates for converging, non-singular multivariate models fitted to blue tit out-of-sample-prediction estimates y_i.*

	Parameter	Random Effect	Coefficient	SE	95%CI	t	df	*p*
blue tit								
y_{50}	(Intercept)		-0.42	0.49	[-1.39,0.56]	-0.84	216	0.4
	publishable_as_ispublishable as is		-0.59	0.43	[-1.44,0.26]	-1.37	216	0.2
	publishable_as_ispublishable with major revision		-0.27	0.33	[-0.92,0.38]	-0.81	216	0.4
	publishable_as_ispublishable with minor revision		-0.4	0.38	[-1.16,0.35]	-1.05	216	0.3
	rate_analysis		3.62×10^{-3}	4.94×10^{-3}	$[-6.13 \times 10^{-3},0.01]$	0.73	216	0.5
	mean_diversity_index		-1.69	0.66	[-2.98,-0.39]	-2.57	216	0.011
	SD (Intercept)	Reviewer ID	0.2	0.09	[0.08,0.49]			
	SD (Observations)	Residual	0.71	0.04	[0.64,0.79]			
y_{75}	(Intercept)		-1.57	0.48	[-2.51,-0.63]	-3.3	229	0.001
	publishable_as_ispublishable as is		0.41	0.41	[-0.4,1.22]	1.01	229	0.3
	publishable_as_ispublishable with major revision		0.1	0.31	[-0.51,0.71]	0.32	229	0.7
	publishable_as_ispublishable with minor revision		0.37	0.37	[-0.36,1.1]	1	229	0.3
	rate_analysis		-4.10×10^{-4}	4.70×10^{-3}	$[-9.66 \times 10^{-3},8.84 \times 10^{-3}]$	-0.09	229	>0.9
	mean_diversity_index		0.13	0.63	[-1.12,1.38]	0.21	229	0.8
	SD (Intercept)	Reviewer ID	0.29	0.07	[0.18,0.45]			
	SD (Observations)	Residual	0.64	0.03	[0.57,0.71]			
Eucalyptus								
y_{50}	(Intercept)		-2.08	1.21	[-4.47,0.31]	-1.72	162	0.087
	publishable_as_ispublishable as is		-0.05	0.86	[-1.75,1.66]	-0.06	162	>0.9
	publishable_as_ispublishable with major revision		0.12	0.58	[-1.03,1.26]	0.2	162	0.8
	publishable_as_ispublishable with minor revision		0.16	0.67	[-1.16,1.48]	0.24	162	0.8
	rate_analysis		4.04×10^{-4}	0.01	[-0.02,0.02]	0.04	162	>0.9
	mean_diversity_index		1.99	1.58	[-1.13,5.1]	1.26	162	0.2
	mixed_model		0.31	0.26	[-0.21,0.83]	1.19	162	0.2
	SD (Intercept)	Reviewer ID	0.09	0.61	$[1.68 \times 10^{-7},4.89 \times 10^{4}]$			
	SD (Observations)	Residual	1.3	0.08	[1.15,1.46]			
y_{75}	(Intercept)		-0.54	1.01	[-2.53,1.44]	-0.54	161	0.6
	publishable_as_ispublishable as is		0.19	0.81	[-1.42,1.79]	0.23	161	0.8
	publishable_as_ispublishable with major revision		-0.01	0.55	[-1.11,1.08]	-0.02	161	>0.9

publishable_as_ispublishable with minor revision		0.17	0.63	[-1.08,1.41]	0.26	161	0.8
rate_analysis		-1.91 × 10-³	9.63 × 10-³	[-0.02,0.02]	-0.2	161	0.8
mean_diversity_index		0.31	1.31	[-2.27,2.9]	0.24	161	0.8
mixed_model		-0.11	0.25	[-0.6,0.38]	-0.46	161	0.6
SD (Intercept)	Reviewer ID	0.04	1.3	[1.69 × 10⁻³⁰,9.00 × 10²⁶]			
SD (Observations)	Residual	1.26	0.08	[1.11,1.42]			

Table A.15: *Model summary statistics for non-singular, converging multivariate models fit to out-of-sample estimates y_i.*

	RMSE	σ	$R^2_{\text{Conditional}}$	R^2_{Marginal}	ICC	N_{Obs}
blue tit						
y_{50}	0.68	0.71	0.11	0.05	0.07	224
y_{75}	0.59	0.64	0.2	0.03	0.17	237
Eucalyptus						
y_{50}	1.27	1.3	0.02	0.02	0	171
y_{75}	1.23	1.26	0.01	0.01	0	170

Table A.16: *Model summary metrics for models of Box-Cox transformed deviation from the mean y_i estimate as a function of categorical peer-review rating, continuous peer-review rating, and Sorensen's index for blue tit and* Eucalyptus *analyses, and also for the inclusion of random effects for* Eucalyptus *analyses. Coefficient of determination, R^2, is reported for models of deviation as a function of Sorensen diversity scores and presence of random effects, while $R^2_{Conditional}$, $R^2_{Marginal}$ and the intra-class correlation (ICC) are reported for models of deviation as explained by peer-review ratings. For all models the residual standard deviation σ, root mean squared error (RMSE) were calculated. The number of observations ($N_{Obs.}$) is displayed for reference.*

	Prediction Scenario	$N_{\text{Obs.}}$	RMSE	σ	R^2	$R^2_{\text{Conditional}}$	R^2_{Marginal}	ICC
Deviation explained by Sorensen's index								
blue tit	y_{25}	63	0.58	0.59	7.76×10^{-4}			
	y_{50}	60	0.72	0.73	0.03			
	y_{75}	63	0.71	0.73	5.99×10^{-3}			
Eucalyptus	y_{25}	38	0.91	0.94	1.13×10^{-3}			
	y_{50}	38	1.29	1.33	4.82×10^{-3}			
	y_{75}	38	1.26	1.29	4.32×10^{-4}			
Deviation explained by continuous ratings								
blue tit	y_{25}	237	0.58	0.58			2.48×10^{-4}	
	y_{50}	224	0.68	0.71		0.1	1.74×10^{-3}	0.1
	y_{75}	237	0.59	0.63		0.2	1.86×10^{-2}	0.18
Eucalyptus	y_{25}	174	0.92	0.93		0.04	1.79×10^{-2}	0.02
	y_{50}	171	1.28	1.29			5.75×10^{-4}	
	y_{75}	170	1.23	1.24			1.44×10^{-4}	
Deviation explained by categorical ratings								
blue tit	y_{25}	237	0.58	0.58			4.35×10^{-3}	
	y_{50}	224	0.68	0.71		0.1	1.3×10^{-2}	0.09
	y_{75}	237	0.59	0.63		0.2	3.36×10^{-2}	0.17
Eucalyptus	y_{25}	174	0.89	0.92		0.07	1.1×10^{-2}	0.06
	y_{50}	171	1.27	1.29		0.01	3.39×10^{-3}	9.14×10^{-3}
	y_{75}	170	1.23	1.25			1.95×10^{-3}	
Deviation explained by inclusion of random effects								
Eucalyptus	y_{25}	38	0.85	0.87	0.14			
	y_{50}	38	1.29	1.33	3.91×10^{-3}			
	y_{75}	38	1.26	1.29	4.76×10^{-3}			

A.3.8 Post-hoc Analysis: checking the use of model weights in all models explaining deviation from the meta-analytic mean

As we describe in Listing A.1, for models of deviation from the meta-analytic mean effect-size, we had intended to use the inverse variance of the Box-Cox transformed deviation from the meta-analytic mean as model weights. Unfortunately using our intended weights specification resulted in invalid transformed response variables for some models whereby extreme outliers were weighted more heavily (two orders of magnitude) than other effect sizes, which caused both issues in the estimated model parameters as well as convergence issues, in particular for models analysing the effect of categorical peer-review rating on deviation from the meta-analytic mean.

The variance of the Box-Cox transformation scores is calculated using the delta method (Equation A.2).

$$\begin{aligned} \mu_{\text{folded}} &= \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left(1 - 2 \times \Phi\left(-\frac{\mu}{\sigma}; 0, 1\right)\right) \\ \text{SE}_{\text{folded}} &= \sqrt{\mu^2 + \sigma^2 - \mu_{\text{folded}}^2} \\ \text{VAR}_{\text{folded}} &= \text{SE}_{\text{folded}}^2 \\ \text{VAR}_{\text{Box-Cox}} &= \text{VAR}_{\text{folded}} \times \left(\lambda \mu_{\text{folded}}^{\lambda-1}\right)^2 \end{aligned} \tag{A.2}$$

Where:

- μ_{folded} is the folded mean of the deviation scores from the mean for effect i,
- $\text{VAR}_{\text{folded}}$ is the folded variance of the deviation scores from the mean for effect i,
- μ is the deviation score from the mean for effect i, calculated as the difference between effect size i and the mean for all effects: $\bar{Z}_r - Z_{r_i}$,
- σ is the square root of the variance of effect i, VZ_r,
- λ is the Box-Cox transformation parameter (Figure A.10), and
- Φ is the standard normal cumulative distribution function.

Which is executed in the `variance_box_cox` function from the the *ManyEcoEvo* package, illustrated in the following code snippet:

```r
variance_box_cox <- function(folded_mu, folded_v, lambda){
# delta method
  variance_bc <- folded_v * (lambda * folded_mu^(lambda - 1))^2
  return(variance_bc)
}

folded_params <- function(abs_dev_score, VZr){
  mu <- abs_dev_score
  sigma <- sqrt(VZr)
  fold_mu <- sigma * sqrt(2/pi) * exp((-mu^2)/(2 * sigma^2)) +
    mu * (1 - 2 * pnorm(-mu/sigma)) # folded abs_dev_score
  fold_se <- sqrt(mu^2 + sigma^2 - fold_mu^2)
  fold_v <- fold_se^2 # folded VZr
  return(list(fold_mu = fold_mu, fold_v = fold_v))
}
```

Listing A.1: Function to calculate the variance of the Box-Cox transformed deviation scores.

We systematically investigated the impact of using different weighting schemes (no weights, inverse-Box-Cox transformed variance-, and inverse folded variance-, of the absolute deviation scores) on model convergence, singularity and other model checking metrics to aid decision-making about the appropriate weighting scheme and random-effects structure for these models. Given the convergence issues we encountered when using the intended weights, and that the desired random effects structure could not be fitted, we also investigated the impact of using different random-effects structures (effect ID, reviewer ID, or both effect ID and reviewer ID) on model convergence and singularity. For each weighting scheme, we fitted models with each different random effects structure to both the blue tit and *Eucalyptus* analyst data, and evaluated model convergence, singularity and model performance using the `performance::compare_performance` function.

A.3.9 Model Weight Investigation Findings

For the blue tit models of deviation influenced by categorical peer-review rating, no models including study ID as a random-effect were able to be properly fitted, across all model weight specifications Table A.17. For *Eucalyptus* models with either no weights, or inverse folded variance as weights, only models with Reviewer ID as the random-effect fitted properly. While models with either Reviewer ID or Study ID as the random-effect passed singularity and convergence fit-checks, and had estimable parameter means when the Inverse Box-Cox variance was used as a model weight Table A.17.

Table A.17: *Singularity and convergence checks for all combinations of model weights and random-effects structure in models of the effect of categorical peer rating on deviation from the analytic mean. For some models, mean estimates of parameter levels for peer-review rating were not able to be estimated.*

Model Weight	Random Effects	Singular Fit?	Model converged?	Means Estimable?
blue tit				
no_weights	RE_rev	no	yes	yes
	RE_study	no	**no**	**no**
	RE_both	**yes**	yes	**no**
inv_bc_var	RE_rev	no	yes	yes
	RE_study	no	yes	**no**
	RE_both	**yes**	yes	**no**
inv_folded_v	RE_rev	no	yes	yes
	RE_study	no	yes	**no**
	RE_both	**yes**	yes	**no**
Eucalyptus				
no_weights	RE_rev	no	yes	yes
	RE_study	no	**no**	**no**
	RE_both	no	yes	**no**
inv_bc_var	RE_rev	no	yes	yes
	RE_study	no	yes	yes
	RE_both	**yes**	yes	**no**
inv_folded_v	RE_rev	no	yes	yes
	RE_study	no	yes	**no**
	RE_both	**yes**	yes	**no**

To check that the alternative weighting methods generated sensible parameter estimates, we generated marginal effects plots for all models that passed the convergence and singularity checks where marginal effects were estimable for both blue tit Figure A.16 and *Eucalyptus* datasets Figure A.15.

Using the inverse Box-Cox transformed variance for model weights resulted in the marginal mean being pulled towards zero for both datasets, however, this was quite extreme for the *Eucalyptus* dataset Figure A.15*A*. When the random-effect for study ID is substituted in place of the reviewer ID, the model means seem more fitting to the data, but uncertainty of estimates seemed artificially small Figure A.15*B*. For this model, marginal means were all equal in both mean estimate and their standard error when the inverse Box-Cox variance was used as a weight with Study ID as the random effect because the model was nearly unidentifiable Table A.18, so we eliminated this specification from further consideration.

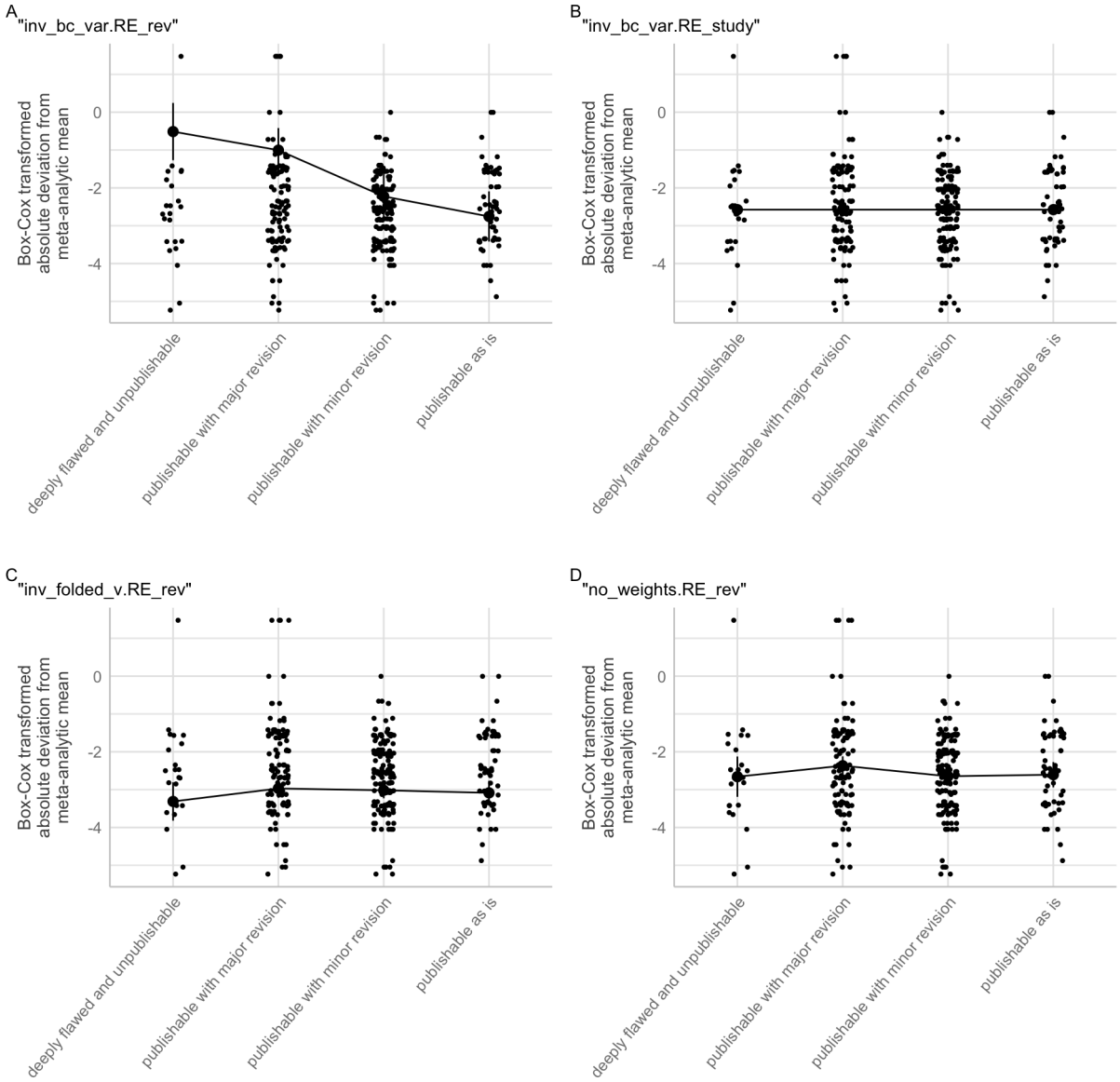


Figure A.15: *Effect plots for each non-singular model that converged with estimable fixed effect group means for the* Eucalyptus *dataset.*

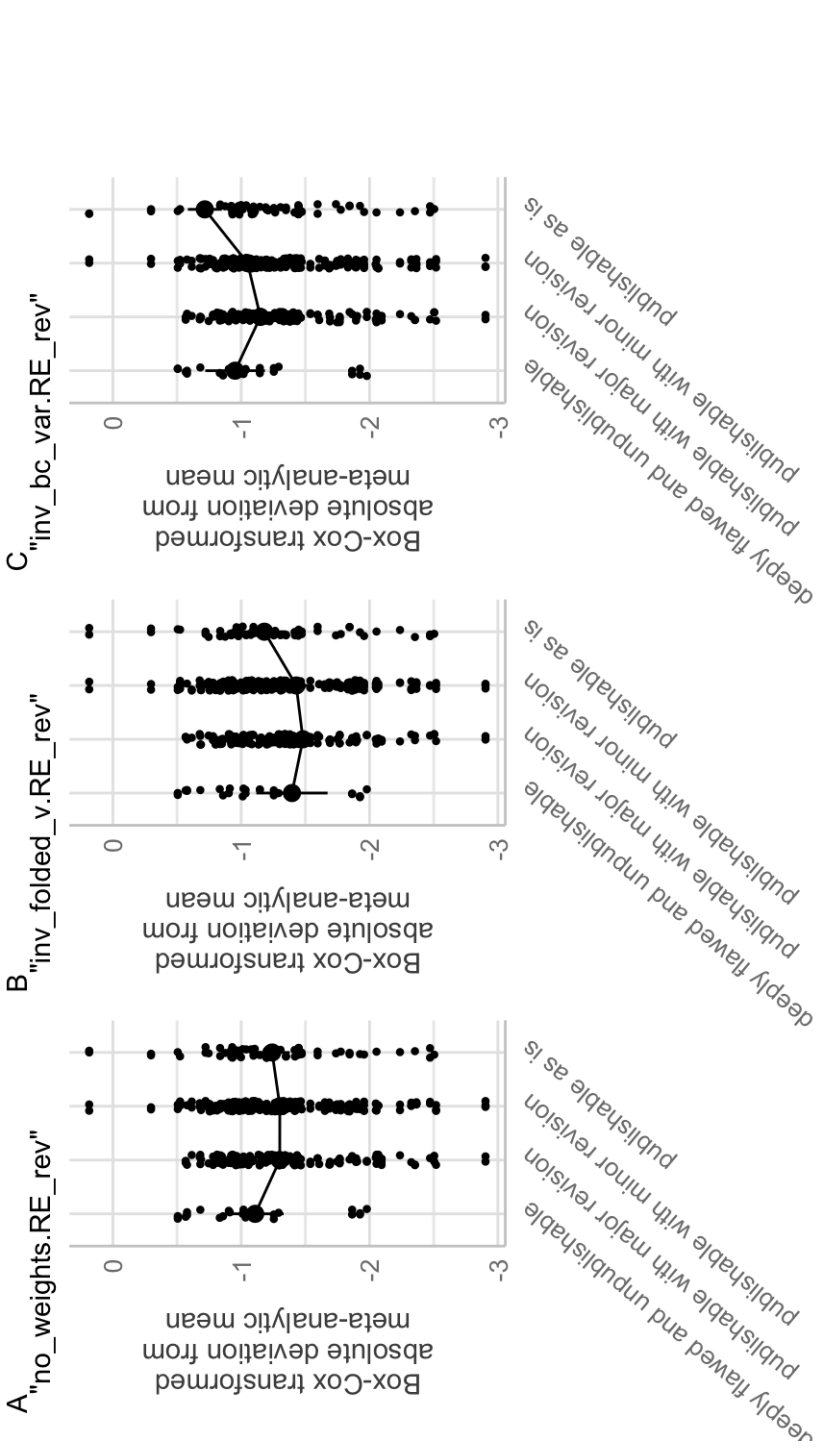
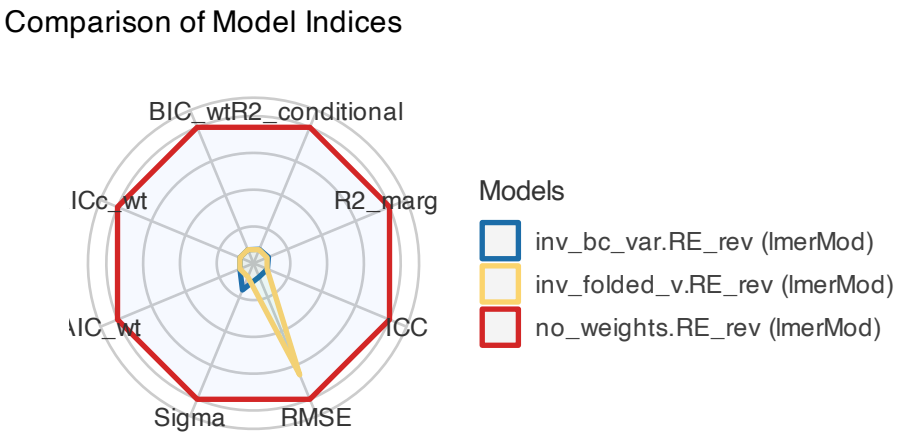


Figure A.16: *Effect plots for each non-singular model that converged with estimable fixed effect group means for the blue tit dataset.*

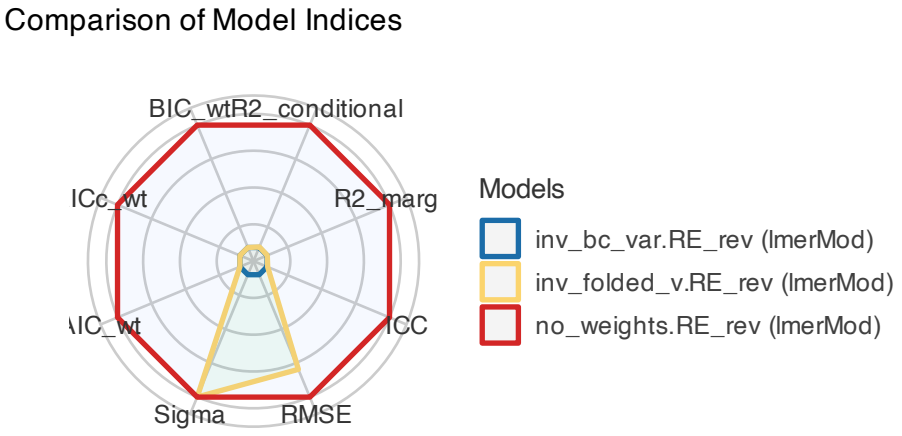
After discarding models based on the above criteria, we were left with a subset of models that passed convergence and singularity checks, and had estimable parameter means, all of which included reviewer ID as the only random effect. Although it seemed that the inverse Box-Cox variance resulted skewed estimated marginal means, it was unclear whether the inverse folded variance was a better alternative over using no weights, as the inverse folded variance seemed to exhibit a similar pattern in pulling the estimated marginal effects towards zero, but to a lesser extent than for the inverse Box-Cox variance.

To further aid in decision-making about which model weights to use, we compared the performance of all models that passed the convergence and singularity checks, and had estimable parameter means. We used the `performance::compare_performance` function to calculate performance metrics for this subset of models, and plotted the results using the `performance` package's in-built plotting method, which creates spider plots of normalised performance metrics for each model Figure A.17. For completeness, all model performance results of this final subset are reported in Table A.19.

The performance comparison plots confirmed our suspicions that the inverse Box-Cox variance was not a suitable weight for these models, and it performed relatively poorly across all metrics for both the *Eucalyptus* and blue tit datasets Figure A.17. The inverse folded variance weighted models performed similarly poorly for both datasets across all metrics except for RMSE in the case of the blue tit dataset, and both Sigma and RMSE for the *Eucalyptus* dataset. For both datasets using no weights when there is only a random-effect for Reviewer ID resulted in the best model fit and performance across all metrics. In keeping with our informed preference of using no weights, model comparison analysis highlighted that using no weights in the models was preferable.



(a) *Blue tit models.*



(b) Eucalyptus *models.*

Figure A.17: *Model performance comparison plots for a subset of models that passed convergence and singularity checks and had estimable marginal effects. Models are compared based on their performance metrics, including* $R^2_{Conditional}$, $R^2_{Marginal}$, *Intra-Class Correlation, Root Mean Squared Error (RMSE), and the weighted AIC, corrected AIC and BIC. Values of performance are normalised across models for each metric, against the top-most performing model across all metrics. Greater distance from the centre on each metric axis indicates dominance in performance. For both blue tit and* Eucalyptus *models, all included a random effect for Reviewer ID, and weights consisted of the inverse Box-Cox variance for weights (blue line), inverse folded variance (yellow line), or none (red line).*

Table A.18: *Marginal means estimated across weight and random effects specifications, for all estimable models, for both* Eucalyptus *and blue tit datasets.*

Model Weight	Random Effects	Peer Rating	Mean	SE	95%CI
blue tit					
no_weights	RE_rev	deeply flawed and unpublishable	-1.11	0.11	[-1.33,-0.89]
		publishable with major revision	-1.30	0.05	[-1.39,-1.21]
		publishable with minor revision	-1.30	0.04	[-1.38,-1.22]
		publishable as is	-1.24	0.07	[-1.38,-1.10]
inv_folded_v	RE_rev	deeply flawed and unpublishable	-1.40	0.14	[-1.67,-1.12]
		publishable with major revision	-1.48	0.06	[-1.61,-1.35]
		publishable with minor revision	-1.43	0.05	[-1.53,-1.32]
		publishable as is	-1.18	0.08	[-1.35,-1.01]
inv_bc_var	RE_rev	deeply flawed and unpublishable	-0.95	0.12	[-1.18,-0.72]
		publishable with major revision	-1.15	0.06	[-1.26,-1.03]
		publishable with minor revision	-1.05	0.05	[-1.15,-0.95]
		publishable as is	-0.72	0.07	[-0.85,-0.58]
Eucalyptus					
no_weights	RE_rev	deeply flawed and unpublishable	-2.66	0.27	[-3.19,-2.12]
		publishable with major revision	-2.37	0.12	[-2.61,-2.12]
		publishable with minor revision	-2.65	0.11	[-2.86,-2.43]
		publishable as is	-2.60	0.17	[-2.94,-2.27]
inv_folded_v	RE_rev	deeply flawed and unpublishable	-3.31	0.26	[-3.82,-2.81]
		publishable with major revision	-2.97	0.13	[-3.21,-2.72]
		publishable with minor revision	-3.02	0.11	[-3.23,-2.80]
		publishable as is	-3.08	0.16	[-3.40,-2.77]
inv_bc_var	RE_rev	deeply flawed and unpublishable	-0.51	0.39	[-1.27,0.24]
		publishable with major revision	-1.00	0.30	[-1.59,-0.42]
		publishable with minor revision	-2.22	0.30	[-2.81,-1.64]
		publishable as is	-2.76	0.34	[-3.42,-2.09]
inv_bc_var	RE_study	deeply flawed and unpublishable	-2.58	0.06	[-2.70,-2.45]
		publishable with major revision	-2.58	0.06	[-2.70,-2.45]
		publishable with minor revision	-2.58	0.06	[-2.70,-2.45]
		publishable as is	-2.58	0.06	[-2.70,-2.45]

Table A.19: *Model performance metric values (non-normalised) for final subset of models considered in weights analysis. All models in final subset included random-effect of Reviewer ID. Metrics included $R^2_{Conditional}$, $R^2_{Marginal}$, Intra-Class Correlation, Root Mean Squared Error (RMSE), and the weighted AIC, corrected AIC and BIC.*

Model Weight	AIC	AIC (wt)	AIC_c	AIC_c (wt)	BIC	BIC (wt)	$R^2_{\text{Conditional}}$	R^2_{Marginal}	ICC	RMSE	σ
blue tit											
None	723.15	1	723.33	1	748.1	1	0.09	7.47×10^{-3}	0.08	0.4827981	0.5
Inverse Box-Cox variance	840.52	3.26×10^{-26}	840.7	3.26×10^{-26}	865.48	3.26×10^{-26}	7.67×10^{-4}	1.31×10^{-4}	6.36×10^{-4}	0.5822505	11.83
Inverse folded variance	1,026.75	1.19×10^{-66}	1,026.93	1.19×10^{-66}	1,051.7	1.19×10^{-66}	4.71×10^{-4}	5.14×10^{-5}	4.2×10^{-4}	0.5026704	13.19
Eucalyptus											
None	1,063.74	1	1,063.99	1	1,086.82	1	0.13	0.01	0.12	1.0173409	1.06
Inverse Box-Cox variance	1,386.88	6.80×10^{-71}	1,387.12	6.80×10^{-71}	1,409.96	6.80×10^{-71}	2.28×10^{-9}	2.47×10^{-10}	2.03×10^{-9}	1.4993872	4.68×10^{4}
Inverse folded variance	1,109.35	1.25×10^{-10}	1,109.6	1.25×10^{-10}	1,132.43	1.25×10^{-10}	5.66×10^{-4}	1.82×10^{-5}	5.48×10^{-4}	1.1266586	19.68

A.4 Correlation Matrices of Case Study Data

Omitted due to page constraints. Please see egouldo.github.io/ManyAnalysts/supp_mat/SM4_case_study_datasets.html in the online quarto manuscript.

A.5 R Package References and Session Info

Table A.20: *R packages used to generate this manuscript. Please see the* ManyEcoEvo *package for a full list of packages used in the analysis pipeline.*

Package	Version	Citation
base	4.4.0	R Core Team (2024)
betapart	1.6	Baselga et al. (2023)
broom.mixed	0.2.9.5	Bolker et al. (2024)
colorspace	2.1.0	Zeileis et al. (2020)
cowplot	1.1.3	Wilke (2024)
devtools	2.4.5	Wickham et al. (2022)
EnvStats	2.8.1	Millard (2013)
GGally	2.2.1	Schloerke et al. (2024)
ggforestplot	0.1.0	Scheinin et al. (2020)
ggh4x	0.2.8	van den Brand (2024)
ggpubr	0.6.0	Kassambara (2023)
ggrepel	0.9.5	Slowikowski (2024)
ggthemes	5.1.0	J. B. Arnold (2024)
glmmTMB	1.1.8	Brooks et al. (2017)
gt	0.10.1	Iannone et al. (2024)
gtsummary	1.7.2	Sjoberg et al. (2021)
here	1.0.1	Müller (2020)
Hmisc	5.1.2	Harrell Jr (2024)
irr	0.84.1	Gamer et al. (2019)
janitor	2.2.0	Firke (2023)
knitr	1.46	Xie (2024a)
latex2exp	0.9.6	Meschiari (2022)
lme4	1.1.35.3	Bates et al. (2015)
ManyEcoEvo	1.1.0	Gould et al. (2023)
metafor	4.6.0	Viechtbauer (2010)
modelbased	0.8.7	Makowski et al. (2025)
multilevelmod	1.0.0	Kuhn & Frick (2022)
MuMIn	1.47.5	Bartoń (2023)
naniar	2.7.6	Tierney & Cook (2023)
NatParksPalettes	0.2.0	Blake (2022)
orchaRd	2.0	Nakagawa, Lagisz, et al. (2023)
parameters	0.21.7	Lüdecke et al. (2020)
patchwork	1.2.0	Pedersen (2024)
performance	0.11.0	Lüdecke, Ben-Shachar, et al. (2021)
renv	1.0.2	Ushey & Wickham (2026)
rmarkdown	2.27	Allaire et al. (2024)
sae	1.3	Molina & Marhuenda (2015)
scales	1.3.0	Wickham, Pedersen, et al. (2023)
see	0.8.4	Lüdecke, Patil, et al. (2021)
showtext	0.9.7	Qiu (2024)
specr	1.0.0	Masur & Scharkow (2020)
targets	1.7.0	Landau (2021)
tidymodels	1.1.1	Kuhn & Wickham (2020)

Table A.20: *R packages used to generate this manuscript. Please see the* ManyEcoEvo *package for a full list of packages used in the analysis pipeline.*

Package	Version	Citation
tidytext	0.4.2	Silge & Robinson (2016)
tidyverse	2.0.0	Wickham et al. (2019)
withr	3.0.0	Hester et al. (2024)
xfun	0.44	Xie (2024b)

Table A.21: *Session information for results generated with the ‘Same Data, Different Analysts’ quarto manuscript (Gould et al., 2024).*

Setting	Value
version	R version 4.4.0 (2024-04-24)
os	macOS Ventura 13.6.9
system	aarch64, darwin20
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Australia/Melbourne
date	2024-09-17
pandoc	3.1.12.2 @ /opt/homebrew/bin/ (via rmarkdown)

B Published Article for the "Same Data, Different Analysts" Study

The following pages include the published article (Gould et al., 2025) presented in Chapter 2, which has been reproduced in entirety with permission under license CC BY-NC-ND 4.0:

> **Gould, E.**, Fraser, H. S., Parker, T. H. et al. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. BMC Biology 23, 35 (2025). https://doi.org/10.1186/s12915-024-02101-x.

Gould *et al. BMC Biology* (2025) 23:35
https://doi.org/10.1186/s12915-024-02101-x

BMC Biology

REGISTERED REPORT

Open Access

Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology

Elliot Gould[1], Hannah S. Fraser[2], Timothy H. Parker[3*], Shinichi Nakagawa[4], Simon C. Griffith[5], Peter A. Vesk[1], Fiona Fidler[2], Daniel G. Hamilton[6], Robin N. Abbey-Lee[7], Jessica K. Abbott[8], Luis A. Aguirre[9], Carles Alcaraz[10], Irith Aloni[11], Drew Altschul[12], Kunal Arekar[13], Jeff W. Atkins[14], Joe Atkinson[15], Christopher M. Baker[16], Meghan Barrett[17], Kristian Bell[18], Suleiman Kehinde Bello[19], Iván Beltrán[20], Bernd J. Berauer[21], Michael Grant Bertram[22], Peter D. Billman[23], Charlie K. Blake[24], Shannon Blake[25], Louis Bliard[26], Andrea Bonisoli-Alquati[27], Timothée Bonnet[28], Camille Nina Marion Bordes[29], Aneesh P. H. Bose[22], Thomas Botterill-James[30], Melissa Anna Boyd[31], Sarah A. Boyle[32], Tom Bradfer-Lawrence[33], Jennifer Bradham[34], Jack A. Brand[22], Martin I. Brengdahl[35], Martin Bulla[36], Luc Bussière[37], Ettore Camerlenghi[38], Sara E. Campbell[39], Leonardo L. F. Campos[40], Anthony Caravaggi[41], Pedro Cardoso[42], Charles J. W. Carroll[43], Therese A. Catanach[44], Xuan Chen[45], Heung Ying Janet Chik[46], Emily Sarah Choy[47], Alec Philip Christie[48], Angela Chuang[49], Amanda J. Chunco[50], Bethany L. Clark[51], Andrea Contina[52], Garth A. Covernton[53], Murray P. Cox[54], Kimberly A. Cressman[55], Marco Crotti[56], Connor Davidson Crouch[57], Pietro B. D'Amelio[58], Alexandra Allison de Sousa[59], Timm Fabian Döbert[60], Ralph Dobler[61], Adam J. Dobson[62], Tim S. Doherty[63], Szymon Marian Drobniak[64], Alexandra Grace Duffy[65], Alison B. Duncan[66], Robert P. Dunn[67], Jamie Dunning[68], Trishna Dutta[69], Luke Eberhart-Hertel[70], Jared Alan Elmore[71], Mahmoud Medhat Elsherif[72], Holly M. English[73], David C. Ensminger[74], Ulrich Rainer Ernst[75], Stephen M. Ferguson[76], Esteban Fernandez-Juricic[77], Thalita Ferreira-Arruda[78], John Fieberg[79], Elizabeth A. Finch[80], Evan A. Fiorenza[81], David N. Fisher[82], Amélie Fontaine[83], Wolfgang Forstmeier[70], Yoan Fourcade[84], Graham S. Frank[85], Cathryn A. Freund[86], Eduardo Fuentes-Lillo[87], Sara L. Gandy[88], Dustin G. Gannon[89], Ana I. García-Cervigón[90], Alexis C. Garretson[91], Xuezhen Ge[92], William L. Geary[93], Charly Géron[94], Marc Gilles[95], Antje Girndt[96], Daniel Gliksman[97], Harrison B. Goldspiel[98], Dylan G. E. Gomes[99], Megan Kate Good[100], Sarah C. Goslee[101], J. Stephen Gosnell[102], Eliza M. Grames[103], Paolo Gratton[104], Nicholas M. Grebe[105], Skye M. Greenler[106], Maaike Griffioen[107], Daniel M. Griffith[108], Frances J. Griffith[109], Jake J. Grossman[110], Ali Güncan[111], Stef Haesen[112], James G. Hagan[113], Heather A. Hager[114], Jonathan Philo Harris[115], Natasha Dean Harrison[116], Sarah Syedia Hasnain[117], Justin Chase Havird[118], Andrew J. Heaton[119], María Laura Herrera-Chaustre[120], Tanner J. Howard[1], Bin-Yan Hsu[121], Fabiola Iannarilli[79], Esperanza C. Iranzo[122], Erik N. K. Iverson[123],

*Correspondence:
Timothy H. Parker
parkerth@whitman.edu
Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Saheed Olaide Jimoh[124], Douglas H. Johnson[79], Martin Johnsson[125], Jesse Jorna[126], Tommaso Jucker[127], Martin Jung[128], Ineta Kačergytė[129], Oliver Kaltz[130], Alison Ke[131], Clint D. Kelly[132], Katharine Keogan[133], Friedrich Wolfgang Keppeler[134], Alexander K. Killion[135], Dongmin Kim[136], David P. Kochan[137], Peter Korsten[138], Shan Kothari[139], Jonas Kuppler[140], Jillian M. Kusch[141], Malgorzata Lagisz[142], Kristen Marianne Lalla[83], Daniel J. Larkin[79], Courtney L. Larson[143], Katherine S. Lauck[131], M. Elise Lauterbur[144], Alan Law[145], Don-Jean Léandri-Breton[83], Jonas J. Lembrechts[146], Kiara L'Herpiniere[5], Eva J. P. Lievens[147], Daniela Oliveira de Lima[148], Shane Lindsay[149], Martin Luquet[150], Ross MacLeod[151], Kirsty H. Macphie[152], Kit Magellan[153], Magdalena M. Mair[154], Lisa E. Malm[155], Stefano Mammola[156], Caitlin P. Mandeville[157], Michael Manhart[158], Laura Milena Manrique-Garzon[159], Elina Mäntylä[121], Philippe Marchand[160], Benjamin Michael Marshall[145], Charles A. Martin[161], Dominic Andreas Martin[162], Jake Mitchell Martin[22], April Robin Martinig[163], Erin S. McCallum[22], Mark McCauley[164], Sabrina M. McNew[144], Scott J. Meiners[165], Thomas Merkling[166], Marcus Michelangeli[22], Maria Moiron[167], Bruno Moreira[168], Jennifer Mortensen[169], Benjamin Mos[170], Taofeek Olatunbosun Muraina[171], Penelope Wrenn Murphy[172], Luca Nelli[56], Petri Niemelä[173], Josh Nightingale[174], Gustav Nilsonne[175], Sergio Nolazco[38], Sabine S. Nooten[176], Jessie Lanterman Novotny[177], Agnes Birgitta Olin[178], Chris L. Organ[179], Kate L. Ostevik[180], Facundo Xavier Palacio[181], Matthieu Paquet[129], Darren James Parker[182], David J. Pascall[183], Valerie J. Pasquarella[184], John Harold Paterson[113], Ana Payo-Payo[185], Karen Marie Pedersen[186], Grégoire Perez[187], Kayla I. Perry[188], Patrice Pottier[142], Michael J. Proulx[189], Raphaël Proulx[190], Jessica L Pruett[191], Veronarindra Ramananjato[192], Finaritra Tolotra Randimbiarison[193], Onja H. Razafindratsima[194], Diana J. Rennison[195], Federico Riva[196], Sepand Riyahi[197], Michael James Roast[198], Felipe Pereira Rocha[199], Dominique G. Roche[200], Cristian Román-Palacios[201], Michael S. Rosenberg[202], Jessica Ross[203], Freya E. Rowland[204], Deusdedith Rugemalila[205], Avery L. Russell[206], Suvi Ruuskanen[207], Patrick Saccone[208], Asaf Sadeh[209], Stephen M. Salazar[210], Kris Sales[211], Pablo Salmón[212], Alfredo Sánchez-Tójar[213], Leticia Pereira Santos[214], Francesca Santostefano[215], Hayden T. Schilling[216], Marcus Schmidt[217], Tim Schmoll[213], Adam C. Schneider[218], Allie E. Schrock[219], Julia Schroeder[68], Nicolas Schtickzelle[220], Nick L. Schultz[221], Drew A. Scott[222], Michael Peter Scroggie[223], Julie Teresa Shapiro[224], Nitika Sharma[225], Caroline L. Shearer[219], Diego Simón[226], Michael I. Sitvarin[227], Fabrício Luiz Skupien[228], Heather Lea Slinn[229], Grania Polly Smith[230], Jeremy A. Smith[231], Rahel Sollmann[89], Kaitlin Stack Whitney[232], Shannon Michael Still[233], Erica F. Stuber[234], Guy F. Sutton[235], Ben Swallow[236], Conor Claverie Taff[237], Elina Takola[238], Andrew J. Tanentzap[239], Rocío Tarjuelo[240], Richard J. Telford[241], Christopher J. Thawley[242], Hugo Thierry[243], Jacqueline Thomson[92], Svenja Tidau[244], Emily M. Tompkins[245], Claire Marie Tortorelli[246], Andrew Trlica[247], Biz R. Turnell[248], Lara Urban[249], Stijn Van de Vondel[146], Jessica Eva Megan van der Wal[250], Jens Van Eeckhoven[251], Francis van Oordt[83], K. Michelle Vanderwel[252], Mark C. Vanderwel[253], Karen J. Vanderwolf[254], Juliana Vélez[79], Diana Carolina Vergara-Florez[255], Brian C. Verrelli[146], Marcus Vinícius Vieira[256], Nora Villamil[257], Valerio Vitali[258], Julien Vollering[259], Jeffrey Walker[260], Xanthe J. Walker[261], Jonathan A. Walter[262], Pawel Waryszak[263], Ryan J. Weaver[264], Ronja E. M. Wedegärtner[265], Daniel L. Weller[266], Shannon Whelan[83], Rachel Louise White[267], David William Wolfson[79], Andrew Wood[268], Scott W. Yanco[269], Jian D. L. Yen[223], Casey Youngflesh[270], Giacomo Zilio[271], Cédric Zimmer[272], Gregory Mark Zimmerman[273] and Rachel A. Zitomer[85]

Abstract

Although variation in effect sizes and predicted values among studies of similar phenomena is inevitable, such variation far exceeds what might be produced by sampling error alone. One possible explanation for variation among results is differences among researchers in the decisions they make regarding statistical analyses. A growing array of studies has explored this analytical variability in different fields and has found substantial variability among results despite analysts having the same data and research question. Many of these studies have been in the social sciences, but one small "many analyst" study found similar variability in ecology. We expanded the scope of this prior work by implementing a large-scale empirical exploration of the variation in effect sizes and model predictions generated by the analytical decisions of different researchers in ecology and evolutionary biology. We used two unpublished datasets, one from evolutionary ecology (blue tit, *Cyanistes caeruleus*, to compare sibling number and nestling growth) and one from conservation ecology (*Eucalyptus*, to compare grass cover and tree seedling recruitment). The project leaders recruited 174 analyst teams, comprising 246 analysts, to investigate the answers to prespecified research questions. Analyses conducted by these teams yielded 141 usable effects (compatible with our meta-analyses and with all necessary information provided) for the blue tit dataset, and 85 usable effects for the *Eucalyptus* dataset. We found substantial heterogeneity among results for both datasets, although the patterns of variation differed between them. For the blue tit analyses, the average effect was convincingly negative, with less growth for nestlings living with more siblings, but there was near continuous variation in effect size from large negative effects to effects near zero, and even effects crossing the traditional threshold of statistical significance in the opposite direction. In contrast, the average relationship between grass cover and *Eucalyptus* seedling number was only slightly negative and not convincingly different from zero, and most effects ranged from weakly negative to weakly positive, with about a third of effects crossing the traditional threshold of significance in one direction or the other. However, there were also several striking outliers in the *Eucalyptus* dataset, with effects far from zero. For both datasets, we found substantial variation in the variable selection and random effects structures among analyses, as well as in the ratings of the analytical methods by peer reviewers, but we found no strong relationship between any of these and deviation from the meta-analytic mean. In other words, analyses with results that were far from the mean were no more or less likely to have dissimilar variable sets, use random effects in their models, or receive poor peer reviews than those analyses that found results that were close to the mean. The existence of substantial variability among analysis outcomes raises important questions about how ecologists and evolutionary biologists should interpret published results, and how they should conduct analyses in the future.

Keywords Analytical heterogeneity, Metascience, Many-analyst, Replication crisis, Reproducibility

Introduction

One value of science derives from its production of replicable, and thus reliable, results. When we repeat a study using the original methods, we should be able to expect a similar result. However, perfect replicability is not a reasonable goal. Effect sizes will vary, and even reverse in sign, by chance alone [37]. Observed patterns can differ for other reasons as well. It could be that we do not sufficiently understand the conditions that led to the original result so when we seek to replicate it, the conditions differ due to some "hidden moderator". This hidden moderator hypothesis is described by meta-analysts in ecology and evolutionary biology as "true biological heterogeneity" [93]. This idea of true heterogeneity is popular in ecology and evolutionary biology, and there are good reasons to expect it in the complex systems in which we work [94]. However, despite similar expectations in psychology, recent evidence in that discipline contradicts the hypothesis that moderators are common obstacles to replicability, as variability in results in a large "many labs" collaboration was mostly unrelated to commonly hypothesized moderators such as the conditions under which the studies were administered [50]. Another possible explanation for variation in effect sizes is that researchers often present biased samples of results, thus reducing the likelihood that later studies will produce similar effect sizes [33, 34, 80, 82, 83]. It also may be that although researchers did successfully replicate the conditions, the experiment, and measured variables, analytical decisions differed sufficiently among studies to create divergent results [96, 99].

Analytical decisions vary among studies because researchers have many options. Researchers need to decide how to exclude possibly anomalous or unreliable data, how to construct variables, which variables to include in their models, and which statistical methods to use. Depending on the dataset, this short list of choices could encompass thousands or millions of possible alternative specifications [100]. However, researchers making these decisions presumably do so with the goal of doing the best possible analysis, or at least the best analysis within their current skill set. Thus, it seems

likely that some specification options are more probable than others, possibly because they have previously been shown (or claimed) to be better, or because they are more well known. Of course, some of these different analyses (maybe many of them) may be equally valid alternatives. Regardless, on probably any topic in ecology and evolutionary biology, we can encounter differences in choices of data analysis. The extent of these differences in analyses and the degree to which these differences influence the outcomes of analyses and therefore studies' conclusions are important empirical questions. These questions are especially important given that many papers draw conclusions after applying a single method, or even a single statistical model, to analyze a dataset.

The possibility that different analytical choices could lead to different outcomes has long been recognized [36], and various efforts to address this possibility have been pursued in the literature. For instance, one common method in ecology and evolutionary biology involves creating a set of candidate models, each consisting of a different (though often similar) set of predictor variables, and then, for the predictor variable of interest, averaging the slope across all models (i.e. model averaging) [20, 40]. This method reduces the chance that a conclusion is contingent upon a single model specification, though use and interpretation of this method is not without challenges [40]. Further, the models compared to each other typically differ only in the inclusion or exclusion of certain predictor variables and not in other important ways, such as methods of parameter estimation. More explicit examination of outcomes of differences in model structure, model type, data exclusion, or other analytical choices can be implemented through sensitivity analyses (e.g., [78]). Sensitivity analyses, however, are typically rather narrow in scope and are designed to assess the sensitivity of analytical outcomes to a particular analytical choice rather than to a large universe of choices. Recently, however, analysts in the social sciences have proposed extremely thorough sensitivity analysis, including 'multiverse analysis' [104] and the 'specification curve' [99], as a means of increasing the reliability of results. With these methods, researchers identify relevant decision points encountered during analysis and conduct the analysis many times to incorporate many plausible decisions made at each of these points. The study's conclusions are then based on a broad set of the possible analyses and so allow the analyst to distinguish between robust conclusions and those that are highly contingent on particular model specifications. These are useful outcomes, but specifying a universe of possible modelling decisions is not a trivial undertaking. Further, the analyst's knowledge and biases will influence decisions about the boundaries of that universe, and so there will always be room for disagreement among analysts about what to include. Including more specifications is not necessarily better. Some analytical decisions are better justified than others, and including biologically implausible specifications may undermine this process. Regardless, these powerful methods have yet to be adopted, and even the more limited forms of sensitivity analyses are not particularly widespread. Most studies publish a small set of analyses and so the existing literature does not provide much insight into the degree to which published results are contingent on analytical decisions.

Despite the potential major impacts of analytical decisions on variance in results, the outcomes of different individuals' data analysis choices have only recently begun to receive much empirical attention. The only formal exploration of this that we were aware of when we submitted our Stage 1 manuscript were (1) an analysis in social science that asked whether male professional football (soccer) players with darker skin tone were more likely to be issued red cards (ejection from the game for rule violation) than players with lighter skin tone [96] and (2) an analysis in neuroimaging which evaluated nine separate hypotheses involving the neurological responses detected with fMRI in 108 participants divided between two treatments in a decision making task [15]. Several others have been published since [16, 23, 44, 92], and we recently learned of an earlier small study in ecology [103]. In the red card study, 29 teams designed and implemented analyses of a dataset provided by the study coordinators [96]. Analyses were peer reviewed (results blind) by at least two other participating analysts,a level of scrutiny consistent with standard pre-publication peer review. Among the final 29 analyses, odds ratios varied from 0.89 to 2.93, meaning point estimates varied from having players with lighter skin tones receive more red cards (odds ratio < 1) to a strong effect of players with darker skin tones receiving more red cards (odds ratio > 1). Twenty of the 29 teams found a statistically significant effect in the predicted direction of players with darker skin tones being issued more red cards. This degree of variation in peer-reviewed analyses from identical data is striking, but the generality of this finding has only just begun to be formally investigated [16, 23, 44, 92].

In the neuroimaging study, 70 teams evaluated each of the nine different hypotheses with the available fMRI data [15]. These 70 teams followed a divergent set of workflows that produced a wide range of results. The rate of reporting of statistically significant support for the nine hypotheses ranged from 21 to 84%, and for each hypothesis on average, 20% of research teams observed effects that differed substantially from the majority of other teams. Some of the variability in results among studies could be explained by analytical decisions such

as choice of software package, smoothing function, and parametric versus non-parametric corrections for multiple comparisons. However, substantial variability among analyses remained unexplained, and presumably emerged from the many different decisions each analyst made in their long workflows. Such variability in results among analyses from this dataset and from the very different red-card dataset suggests that sensitivity of analytical outcome to analytical choices may characterize many distinct fields, as several more recent many-analyst studies also suggest [16, 44, 92].

To further develop the empirical understanding of the effects of analytical decisions on study outcomes, we chose to estimate the extent to which researchers' data analysis choices drive differences in effect sizes, model predictions, and qualitative conclusions in ecology and evolutionary biology. This is an important extension of the meta-research agenda of evaluating factors influencing replicability in ecology, evolutionary biology, and beyond [31]. To examine the effects of analytical decisions, we used two different datasets and recruited researchers to analyze one or the other of these datasets to answer a question we defined. The first question was "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" To answer this question, we provided a dataset that includes brood size manipulations from 332 broods conducted over 3 years at Wytham Wood, UK. The second question was "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" For this question, analysts used a dataset that includes, among other variables, number of seedlings in different size classes, percentage cover of different life forms, tree canopy cover, and distance from canopy edge from 351 quadrats spread among 18 sites in Victoria, Australia.

We explored the impacts of data analysts' choices with descriptive statistics and with a series of tests to attempt to explain the variation among effect sizes and predicted values of the dependent variable produced by the different analysis teams for both datasets separately. To describe the variability, we present forest plots of the standardized effect sizes and predicted values produced by each of the analysis teams, estimate heterogeneity (both absolute,τ^2, and proportional, I^2) in effect size and predicted values among the results produced by these different teams, and calculate a similarity index that quantifies variability among the predictor variables selected for the different statistical models constructed by the different analysis teams. These descriptive statistics provide the first estimates of the extent to which explanatory statistical models and their outcomes in ecology and evolutionary biology vary based on the decisions of different data analysts. We then quantified the degree to which the variability in effect size and predicted values could be explained by (1) variation in the quality of analyses as rated by peer reviewers and (2) the similarity of the choices of predictor variables between individual analyses.

Methods

This project involved a series of steps (1–6) that began with identifying datasets for analyses and continued through recruiting independent groups of scientists to analyze the data, allowing the scientists to analyze the data as they saw fit, generating peer review ratings of the analyses (based on methods, not results), evaluating the variation in effects among the different analyses, and producing the final manuscript.

Step 1: Select datasets

We used two previously unpublished datasets, one from evolutionary ecology and the other from ecology and conservation.

Evolutionary ecology

Our evolutionary ecology dataset is relevant to a subdiscipline of life-history research which focuses on identifying costs and trade-offs associated with different phenotypic conditions. These data were derived from a brood-size manipulation experiment imposed on wild birds nesting in boxes provided by researchers in an intensively studied population. Understanding how the growth of nestlings is influenced by the numbers of siblings in the nest can give researchers insights into factors such as the evolution of clutch size, determination of provisioning rates by parents, and optimal levels of sibling competition [25, 77, 89, 110, 112]. Data analysts were provided this dataset and instructed to answer the following question: "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?"

Researchers conducted brood size manipulations and population monitoring of blue tits at Wytham Wood, a 380-ha woodland in Oxfordshire, UK (1° 20′ W, 51° 47′ N). Researchers regularly checked approximately 1100 artificial nest boxes at the site and monitored the 330 to 450 blue tit pairs occupying those boxes in 2001–2003 during the experiment. Nearly all birds made only one breeding attempt during the April to June study period in a given year. At each blue tit nest, researchers recorded the date the first egg appeared, clutch size, and hatching date. For all chicks alive at age 14 days, researchers measured mass and tarsus length and fitted a uniquely numbered, British Trust for Ornithology (BTO) aluminum leg ring.

Researchers attempted to capture all adults at their nests between day 6 and day 14 of the chick-rearing period. For these captured adults, researchers measured mass, tarsus length, and wing length and fitted a uniquely numbered BTO leg ring. During the 2001–2003 breeding seasons, researchers manipulated brood sizes using cross fostering. They matched broods for hatching date and brood size and moved chicks between these paired nests 1 or 2 days after hatching. They sought to either enlarge or reduce all manipulated broods by approximately one fourth. To control for effects of being moved, each reduced brood had a portion of its brood replaced by chicks from the paired increased brood, and vice versa. Net manipulations varied from plus or minus four chicks in broods of 12 to 16 to plus or minus one chick in broods of 4 or 5. Researchers left approximately one third of all broods unmanipulated. These unmanipulated broods were not selected systematically to match manipulated broods in clutch size or laying date. We have mass and tarsus length data from 3720 individual chicks divided among 167 experimentally enlarged broods, 165 experimentally reduced broods, and 120 unmanipulated broods. The full list of variables included in the dataset is publicly available (https://osf.io/hdv8m), along with the data (https://osf.io/qjzby).

Ecology and conservation

Additional Explanation:

Shortly after beginning to recruit analysts, several analysts noted a small set of related errors in the blue tit dataset. We corrected the errors, replaced the dataset on our OSF site, and emailed the analysts on 19 April 2020 to instruct them to use the revised data. The email to analysts is available here (https://osf.io/4h53z). The errors are explained in that email.

Our ecology and conservation dataset is relevant to a sub-discipline of conservation research which focuses on investigating how best to revegetate private land in agricultural landscapes. These data were collected on private land under the Bush Returns program, an incentive system where participants entered into a contract with the Goulburn Broken Catchment Management Authority and received annual payments if they executed predetermined restoration activities. This particular dataset is based on a passive regeneration initiative, where livestock grazing was removed from the property in the hopes that the *Eucalyptus* spp. overstorey would regenerate without active (and expensive) planting. Analyses of some related data have been published [67, 113] but those analyses do not address the question analysts answered in our study. Data analysts were provided this dataset and instructed to answer the following question: "How does grass cover influence *Eucalyptus* spp. seedling recruitment?".

Researchers conducted three rounds of surveys at 18 sites across the Goulburn Broken catchment in northern Victoria, Australia, in winter and spring 2006 and autumn 2007. In each survey period, a different set of 15 × 15 m quadrats were randomly allocated across each site within 60 m of existing tree canopies. The number of quadrats at each site depended on the size of the site, ranging from four at smaller sites to 11 at larger sites. The total number of quadrats surveyed across all sites and seasons was 351. The number of *Eucalyptus* spp. seedlings was recorded in each quadrat along with information on the GPS location, aspect, tree canopy cover, distance to tree canopy, and position in the landscape. Ground layer plant species composition was recorded in three 0.5 × 0.5 m sub-quadrats within each quadrat. Subjective cover estimates of each species as well as bare ground, litter, rock and moss/lichen/soil crusts were recorded. Subsequently, this was augmented with information about the precipitation and solar radiation at each GPS location. The full list of variables included in the dataset is publicly available (https://osf.io/r5gbn), along with the data (https://osf.io/qz5cu).

Step 2: Recruitment and initial survey of analysts

The lead team (TP, HF, SN, EG, SG, PV, DH, FF) created a publicly available document providing a general description of the project (https://osf.io/mn5aj/). The project was advertised at conferences, via Twitter, using mailing lists for ecological societies (including *Ecolog*, *Evoldir*, and lists for *the Environmental Decisions Group*, and *Transparency in Ecology and Evolution*), and via word of mouth. The target population was active ecology, conservation, or evolutionary biology researchers with a graduate degree (or currently studying for a graduate degree) in a relevant discipline. Researchers could choose to work independently or in a small team. For the sake of simplicity, we refer to these as "analysis teams" though some comprised one individual. We aimed for a minimum of 12 analysis teams independently evaluating each dataset (see sample size justification below). We simultaneously recruited volunteers to peer review the analyses conducted by the other volunteers through the same channels. Our goal was to recruit a similar number of peer reviewers and analysts, and to ask each peer reviewer to review a minimum of four analyses. If we were unable to recruit at least half the number of reviewers as analysis teams, we planned to ask analysts to serve also as reviewers (after they had completed their analyses), but this was unnecessary. Therefore, no data analysts peer reviewed analyses of the dataset they had analyzed. All analysts and reviewers were offered the

opportunity to share co-authorship on this manuscript and we planned to invite them to participate in the collaborative process of producing the final manuscript. All analysts signed [digitally] a consent (ethics) document (https://osf.io/xyp68/) approved by the Whitman College Institutional Review Board prior to being allowed to participate.

Preregistration Deviation:

Due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We identified our minimum number of analysts per dataset by considering the number of effects needed in a meta-analysis to generate an estimate of heterogeneity (τ^2) with a 95% confidence interval that does not encompass zero. This minimum sample size is invariant regardless of τ^2. This is because the same *t*-statistic value will be obtained by the same sample size regardless of variance (τ^2). We see this by first examining the formula for the standard error, SE for variance, (τ^2) or (SEτ^2) assuming normality in an underlying distribution of effect sizes [51]:

$$\mathrm{SE}(\tau^2) = \sqrt{\frac{2\tau^4}{n-1}}$$

and then rearranging the above formula to show how the *t*-statistic is independent of τ^2, as seen below.

$$t = \frac{\tau^2}{\mathrm{SE}(\tau^2)} = \sqrt{\frac{n-1}{2}}$$

We then find a minimum $n = 12$ according to this formula.

Step 3: Primary data analyses

Analysis teams registered and answered a demographic and expertise survey (https://osf.io/seqzy/). We then provided them with the dataset of their choice and requested that they answer a specific research question. For the evolutionary ecology dataset that question was "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" and for the conservation ecology dataset it was "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" Once their analysis was complete, they answered a structured survey (https://osf.io/neyc7/), providing analysis technique, explanations of their analytical choices, quantitative results, and a statement describing their conclusions. They also were asked to upload their analysis files (including the dataset as they formatted it for analysis and their analysis code [if applicable]) and a detailed journal-ready statistical methods section.

Additional Information:

As is common in many studies in ecology and evolutionary biology, the datasets we provided contained many variables, and the research questions we provided could be addressed by our datasets in many different ways. For instance, volunteer analysts had to choose the dependent (response) variable and the independent variable, and make numerous other decisions about which variables and data to use and how to structure their model.

Preregistration Deviation:

We originally planned to have analysts complete a single survey (https://osf.io/neyc7/), but after we evaluated the results of that survey, we realized we would need a second survey (https://osf.io/8w3v5/) to adequately collect the information we needed to evaluate heterogeneity of results (step 5). We provided a set of detailed instructions with the follow-up survey, and these instructions are publicly available and can be found within the following files (blue tit: https://osf.io/kr2g9, *Eucalyptus*: https://osf.io/dfvym).

Step 4: Peer reviews of analyses

At minimum, each analysis was evaluated by four different reviewers, and each volunteer peer reviewer was randomly assigned methods sections from at least four analyst teams (the exact number varied). Each peer reviewer registered and answered a demographic and expertise survey identical to that asked of the analysts, except we did not ask about "team name" since reviewers did not work in teams. Reviewers evaluated the methods of each of their assigned analyses one at a time in a sequence determined by the project leaders. We systematically assigned the sequence so that, if possible, each analysis was allocated to each position in the sequence for at least one reviewer. For instance, if each reviewer were assigned four analyses to review, then each analysis would be the first analysis assigned to at least one reviewer, the second analysis assigned to another reviewer, the third analysis assigned to yet another reviewer, and the fourth analysis assigned to a fourth reviewer. Balancing the order in which reviewers saw the analyses controls for order effects, e.g. a reviewer might be less critical of the first methods section they read than the last.

The process for a single reviewer was as follows. First, the reviewer received a description of the methods of a single analysis. This included the narrative methods section, the analysis team's answers to our survey questions regarding their methods, including analysis code, and the dataset. The reviewer was then asked, in an online survey (https://osf.io/4t36u/), to rate that analysis on a scale of 0–100 based on this prompt: "Rate the overall appropriateness of this analysis to answer

the research question (*one of the two research questions inserted here*) with the available data". To help you calibrate your rating, please consider the following guidelines:

- 100. A perfect analysis with no conceivable improvements from the reviewer
- 75. An imperfect analysis but the needed changes are unlikely to dramatically alter outcomes
- 50. A flawed analysis likely to produce either an unreliable estimate of the relationship or an over-precise estimate of uncertainty
- 25. A flawed analysis likely to produce an unreliable estimate of the relationship and an over-precise estimate of uncertainty
- 0. A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise estimate of uncertainty that places undue confidence in the incorrect estimate.

*Please note that these values are meant to calibrate your ratings. We welcome ratings of any number between 0 and 100.

After providing this rating, the reviewer was presented with this prompt, in multiple-choice format: "Would the analytical methods presented produce an analysis that is (a) publishable as is, (b) publishable with minor revision, (c) publishable with major revision, (d) deeply flawed and unpublishable?" The reviewer was then provided with a series of text boxes and the following prompts: "Please explain your ratings of this analysis. Please evaluate the choice of statistical analysis type. Please evaluate the process of choosing variables for and structuring the statistical model. Please evaluate the suitability of the variables included in (or excluded from) the statistical model. Please evaluate the suitability of the structure of the statistical model. Please evaluate choices to exclude or not exclude subsets of the data. Please evaluate any choices to transform data (or, if there were no transformations, but you think there should have been, please discuss that choice)." After submitting this review, a methods section from a second analysis was then made available to the reviewer. This same sequence was followed until all analyses allocated to a given reviewer were provided and reviewed. After providing the final review, the reviewer was simultaneously provided with all four (or more) methods sections the reviewer had just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation. The invitation to revise the original ratings was as follows: "If, now that you have seen all the analyses you are reviewing, you wish to revise your ratings of any of these analyses, you may do so now." The text box was prefaced with this prompt: "Please explain your choice to revise (or not to revise) your ratings."

Additional Information: unregistered analysis

To determine how consistent peer reviewers were in their ratings, we assessed inter-rater reliability among reviewers for both the categorical and quantitative ratings combining blue tit and *Eucalyptus* data using Krippendorff's alpha for ordinal and continuous data respectively. This provides a value that is between -1 (total disagreement between reviewers) and 1 (total agreement between reviewers).

Step 5: Evaluate variation

Additional Information: analysis schematic

The lead team conducted a range of preregistered and exploratory analyses to understand variation between analyses and their results. Figure 1 is intended to clarify the analyses described below.

The lead team conducted the analyses outlined in this section. We described the variation in model specification in several ways. We calculated summary statistics describing variation among analyses, including mean, SD, and range of number of variables per model included as fixed effects, the number of interaction terms, the number of random effects, and the mean, SD, and range of sample sizes. We also present the number of analyses in which each variable was included. We summarized the variability in standardized effect sizes and predicted values of dependent variables among the individual analyses using standard random effects meta-analytic techniques. First, we derived standardized effect sizes from each individual analysis. We did this for all linear models or generalized linear models by converting the t value and the degree of freedom (df) associated with regression coefficients (e.g. the effect of the number of siblings [predictor] on growth [response] or the effect of grass cover [predictor] on seedling recruitment [response]) to the correlation coefficient, r, using the following:

$$r = \frac{t^2}{(t^2 + df)}$$

This formula can only be applied if t and df values originate from linear or generalized linear models [72]. If, instead, linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) were used by a given analysis, the exact df cannot be estimated. However, adjusted df can be estimated,

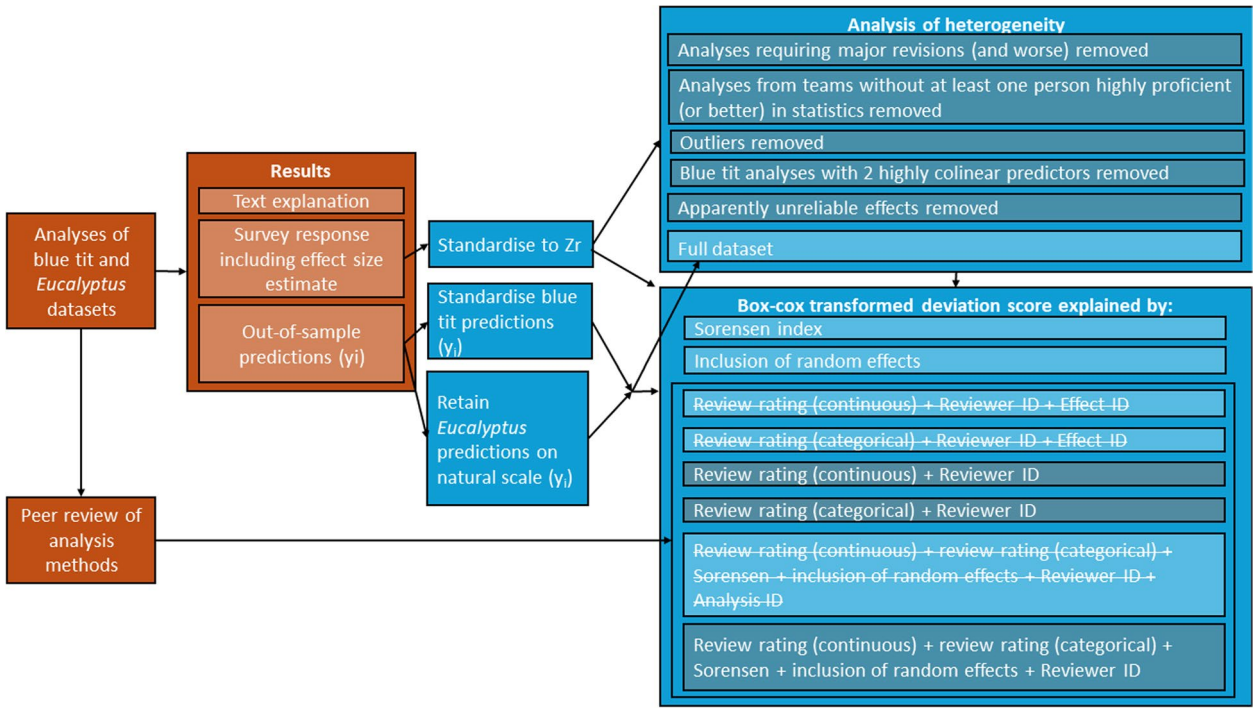


Fig. 1 Schematic of research process showing recruited analyst and reviewer contributions in orange and core team contributions in blue. Items that are crossed out were preregistered but could not be conducted. Items with a greyed background were added as exploratory analyses after preregistration

for example, using the Satterthwaite approximation of *df*, df_S, [note that SAS uses this approximation to obtain *df* for LMMs and GLMMs; [63]. For analyses using either LMMs or GLMMs that do not produce df_S we planned to obtain df_S by rerunning the same (G)LMMs using the lmer() or glmer() function in the *lmerTest* package in R [56, 87].

Preregistration Deviation:

Rather than re-run these analyses ourselves, we sent a follow-up survey (referenced above under "Primary data analyses") to analysts and asked them to follow our instructions for producing this information. The instructions are publicly available and can be found within the following files (blue tit: https://osf.io/kr2g9, *Eucalyptus*: https://osf.io/dfvym).

We then used the *t* values and df_S from the models to obtain *r* as per the formula above. All *r* and accompanying *df* (or df_S) were converted to Fisher's Z_r.

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

and its sampling variance; $1/(n-3)$ where $n=df+1$. Any analyses from which we could not derive a signed Z_r, for instance one with a quadratic function in which the slope changed sign, were considered unusable for analyses of Z_r. We expected such analyses would be rare. In fact, most submitted analyses excluded from our meta-analysis of Z_r were excluded because of a lack of sufficient information provided by the analyst team rather than due to the use of effects that could not be converted to Z_r. Regardless, as we describe below, we generated a second set of standardized effects (predicted values) that could (in principle) be derived from any explanatory model produced by these data.

Besides Z_r, which describes the strength of a relationship based on the amount of variation in a dependent variable explained by variation in an independent variable, we also examined differences in the shape of the relationship between the independent and dependent variables. To accomplish this, we derived a point estimate (out-of-sample predicted value) for the dependent variable of interest for each of three values of our primary independent variable. We originally described these three values as associated with the 25th percentile, median, and 75th percentile of the independent variable and any covariates.

Preregistration Deviation:

The original description of the out-of-sample specifications did not account for the facts that (a) some variables are not distributed in a way that allowed division in percentiles and that (b) variables could be either positively or negatively correlated with the dependent variable. We provide a more thorough description here:

We derived three point-estimates (out-of-sample predicted values) for the dependent variable of interest; one for each of three values of our primary independent variable that we specified. We also specified values for all other variables that could have been included as independent variables in analysts' models so that we could derive the predicted values from a fully specified version of any model produced by analysts. For all potential independent variables, we selected three values or categories. Of the three we selected, one was associated with small, one with intermediate, and one with large values of one typical dependent variable (day 14 chick weight for the blue tit data and total number of seedlings for the *Eucalyptus* data; analysts could select other variables as their dependent variable, but the others typically correlated with the two identified here). For continuous variables, this means we identified the 25th percentile, median, and 75th percentile and, if the slope of the linear relationship between this variable and the typical dependent variable was positive, we left the quartiles ordered as is. If, instead, the slope was negative, we reversed the order of the independent variable quartiles so that the 'lower' quartile value was the one associated with the lower value for the dependent variable. In the case of categorical variables, we identified categories associated with the 25th percentile, median, and 75th percentile values of the typical dependent variable after averaging the values for each category. However, for some continuous and categorical predictors, we also made selections based on the principle of internal consistency between certain related variables, and we fixed a few categorical variables as identical across all three levels where doing so would simplify the modelling process (specification tables available: blue tit: https://osf.io/86akx; *Eucalyptus*: https://osf.io/jh7g5).

We used the 25th and 75th percentiles rather than minimum and maximum values to reduce the chance of occupying unrealistic parameter space. We planned to derive these predicted values from the model information provided by the individual analysts. All values (predictions) were first transformed to the original scale along with their standard errors (SE); we used the delta method [111] for the transformation of SE. We used the square of the SE associated with predicted values as the sampling variance in the meta-analyses described below, and we planned to analyze these predicted values in exactly the same ways as we analyzed Z_r in the following analyses.

Preregistration Deviation:

1. Standardizing blue tit out-of-sample predictions (y_i)

Because analysts of blue tit data chose different dependent variables on different scales, after transforming out-of-sample values to the original scales, we standardized all values as *z* scores ('standard scores') to put all dependent variables on the same scale and make them comparable. This involved taking each relevant value on the original scale (whether a predicted point estimate or a SE associated with that estimate) and subtracting the value in question from the mean value of that dependent variable derived from the full dataset and then dividing this difference by the standard deviation, SD, corresponding to the mean from the full dataset (Supplementary Material B, Equation B.1).

Note that we were unable to standardise some analyst-constructed variables, so these analyses were excluded from the final out-of-sample estimates meta-analysis, see Supplementary Material B, section B.1.2.1 for details and explanation.

2. Log-transforming *Eucalyptus* out-of-sample predictions *yi*

All analyses of the *Eucalyptus* data chose dependent variables that were on the same scale, that is, *Eucalyptus* seedling counts. Although analysts may have used different size-classes of *Eucalyptus* seedlings for their dependent variable, we considered these choices to be akin to subsetting, rather than as different response variables, since changing the size-class of the dependent variable ultimately results in observations being omitted or included. Consequently, we did not standardise *Eucalyptus* out-of-sample predictions.

We were unable to fit quasi-Poisson or Poisson meta-regressions, as desired [79], because available meta-analysis packages (e.g. metafor:: and metainc::) do not provide implementation for outcomes as estimates-only, methods are only provided for outcomes as ratios or rate-differences between two groups. Consequently, we log-transformed the out-of-sample predictions for the *Eucalyptus* data and use the mean estimate for each prediction scenario as the dependent variable in our meta-analysis with the associated SE as the sampling variance in the meta-analysis [74, 75]. Table2.

We plotted individual effect size estimates (Z_r) and predicted values of the dependent variable (y_i) and their corresponding 95% confidence / credible intervals in forest plots to allow visualization of the range and precision of effect size and predicted values. Further, we included these estimates in random effects meta-analyses [14, 43] using the *metafor* package in R [87, 114]:

$$Z_r \sim 1 + \left(1|\textit{Effect ID}\right)$$

$$y_i \sim 1 + \left(1|\textit{Effect ID}\right)$$

where y_i is the predicted value for the dependent variable at the 25th percentile, median, or 75th percentile of the independent variables. The individual Z_r effect sizes were weighted with the inverse of sampling variance for Z_r. The individual predicted values for dependent variable (y_i) were weighted by the inverse of the associated SE^2 (original registration omitted "inverse of the" in error). These analyses provided an average Z_r score ($\overline{Z}_r$) or an average y_i ($\overline{y}_i$) with corresponding 95% confidence interval and allowed us to estimate two heterogeneity indices, τ^2 and I^2. The former, τ^2, is the absolute measure of heterogeneity or the between-study variance (in our case, between-effect variance) whereas I^2 is a relative measure of heterogeneity. We obtained the estimate of relative heterogeneity (I^2) by dividing the between-effect variance by the sum of between-effect and within-effect variance (sampling error variance). I^2 is thus, in a standard meta-analysis, the proportion of variance that is due to heterogeneity as opposed to sampling error. When calculating I^2, within-study variance is amalgamated across studies to create a "typical" within-study variance which serves as the sampling error variance [14, 43]. Our goal

here was to visualize and quantify the degree of variation among analyses in effect size estimates [72]. We did not test for statistical significance.

Additional explanation:

Our use of I^2 to quantify heterogeneity violates an important assumption, but this violation does not invalidate our use of I^2 as a metric of how much heterogeneity can derive from analytical decisions. In standard meta-analysis, the statistic I^2 quantifies the proportion of variance that is greater than we would expect if differences among estimates were due to sampling error alone [88]. However, it is clear that this interpretation does not apply to our value of I^2 because I^2 assumes that each estimate is based on an independent sample (although these analyses can account for non-independence via hierarchical modelling), whereas all our effects were derived from largely or entirely overlapping subsets of the same dataset. Despite this, we believe that I^2 remains a useful statistic for our purposes. This is because, in calculating I^2, we are still setting a benchmark of expected variation due to sampling error based on the variance associated with each separate effect size estimate, and we are assessing how much (if at all) the variability among our effect sizes exceeds what would be expected had our effect sizes been based on independent data. In other words, our estimates can tell us how much proportional heterogeneity is possible from analytical decisions alone when sample sizes (and therefore meta-analytic within-estimate variance) are similar to the ones in our analyses. Among other implications, our violation of the independent sample assumption means that we (dramatically) over-estimate the variance expected due to sampling error, and because I^2 is a proportional estimate, we thus underestimate the actual proportion of variance due to differences among analyses other than sampling error. However, correcting this underestimation would create a trivial value since we designed the study so that much of the variance would derive from analytic decisions as opposed to differences in sampled data. Instead, retaining the I^2 value as typically calculated provides a useful comparison to I^2 values from typical meta-analyses.

Interpretation of τ^2 also differs somewhat from traditional meta-analysis, and we discuss this further in the Results.

Finally, we assessed the extent to which deviations from the meta-analytic mean by individual effect sizes (Z_r) or the predicted values of the dependent variable (y_i) were explained by the peer rating of each analysis team's method section, by a measurement of the distinctiveness of the set of predictor variables included in each analysis, and by the choice of whether or not to include random effects in the model. The deviation score, which served as the dependent variable in these analyses, is the absolute value of the difference between the meta-analytic mean $\overline{Z}_r$ (or $\overline{y}_i$) and the individual Z_r (or y_i) estimate for each analysis. We used the Box-Cox transformation on the absolute values of deviation scores to achieve an approximately normal distribution [28, 29]. We described variation in this dependent variable with both a series of univariate analyses and a multivariate analysis. All these analyses were general linear (mixed) models. These analyses were secondary to our estimation of variation in effect sizes described above. We wished to quantify relationships among variables, but we had no a priori expectation of effect size and made no dichotomous decisions about statistical significance.

When examining the extent to which reviewer ratings (on a scale from 0 to 100) explained deviation from the average effect (or predicted value), each analysis had been rated by multiple peer reviewers, so for each reviewer score to be included, we include each deviation score in the analysis multiple times. To account for the non-independence of multiple ratings of the same analysis, we planned to include analysis identity as a random effect in our general linear mixed model in the *lme4* package in R [11, 87]. To account for potential differences among reviewers in their scoring of analyses, we also planned to include reviewer identity as a random effect:

$$\text{DeviationScore}_j = BoxCox\left(\text{DeviationFromMean}_j\right)$$

$$\text{DeviationScore}_{ij} \sim \text{Rating}_{ij} + \text{ReviewerID}_i + EffectID_j$$

$$\text{ReviewerID}_i \sim N(0, \sigma_i^2)$$

$$EffectID \sim N(0, \sigma_j^2)$$

where $\text{DeviationFromMean}_j$ is the deviation from the meta-analytic mean for the jth analysis, Reviewer ID_i is the random intercept assigned to each i reviewer, and Effect ID_j is the random intercept assigned to each j analysis, both of which are assumed to be normally distributed with a mean of 0 and a variance of σ^2. Absolute deviation scores were Box-Cox transformed using the step_box_cox() function from the *timetk* package in R [24, 87].

Additional explanation:

In our meta-analyses based on Box-Cox transformed deviation scores, we leave these deviation scores unweighted. This is consistent with our registration, which did not mention weighting these scores. However, the fact that we did not mention weighting the scores was actually an error: we had intended to weight them, as is standard in meta-analysis, using the inverse variance of the Box-Cox transformed deviation scores Supplementary Material C, equation C.1. Unfortunately, when we did conduct the weighted analyses, they produced results in which some weighted estimates differed radically from the unweighted estimate because the weights were invalid. Such invalid weights can sometimes occur when the variance (upon which the weights depend) is partly a function of the effect size, as in our Box-Cox transformed deviation scores [73]. In the case of the *Eucalyptus* analyses, the most extreme outlier was weighted much more heavily (by close to two orders of magnitude) than any other effect sizes because the effect size was, itself, so high. Therefore, we made the decision to avoid weighting by inverse variance in all analyses of the Box-Cox transformed deviation scores. This was further justified because (a) most analyses have at least some moderately unreliable weights, and (b) the sample sizes were mostly very similar to each other across submitted analyses, and so meta-analytic weights are not particularly important here [19]. We systematically investigated the impact of different weighting schemes and random effects on model convergence and results, see Supplementary Material C, section C.8 for more details.

We conducted a similar analysis with the four categories of reviewer ratings ((1) deeply flawed and unpublishable, (2) publishable with major revision, (3) publishable with minor revision, (4) publishable as is) set as ordinal predictors numbered as shown here. As with the analyses above, we planned for these analyses to also include random effects of analysis identity and reviewer identity. Both of these analyses (1: 1–100 ratings as the fixed effect, 2: categorical ratings as the fixed effects) were planned to be conducted eight times for each dataset. Each of the four responses (Z_r, y_{25}, y_{50}, y_{75}) were to be compared once to the initial ratings provided by the peer reviewers, and again based on the revised ratings provided by the peer reviewers.

Preregistration deviation:

1. We planned to include random effects of both analysis identity and reviewer identity in these models comparing reviewer ratings with deviation scores. However, after we received the analyses, we discovered that a subset of analyst teams had either conducted multiple analyses and/or identified multiple effects per analysis as answering the target question. We therefore faced an even more complex potential set of random effects. We decided that including Team ID and Effect ID along with Reviewer ID as random effects in the same model would almost certainly lead to model fit problems, and so we started with simpler models including just Effect ID and Reviewer ID. However, even with this simpler structure, our dataset was sparse, with reviewers rating a small number of analyses, resulting in models with singular fit (Supplementary Material C, section C.2). Removing one of the random effects was necessary for the models to converge. For both models of deviation from the meta-analytic mean explained by categorical or continuous reviewer ratings, we removed the random effect of Effect ID, leaving Reviewer ID as the only random effect.

2. We conducted analyses only with the final peer ratings after the opportunity for revision, not with the initial ratings. This was because when we recorded the final ratings, the initial ratings were over-written, therefore we did not have access to those initial values.

The next set of univariate analyses sought to explain deviations from the mean effects based on a measure of the distinctiveness of the set of variables included in each analysis. As a 'distinctiveness' score, we used Sorensen's Similarity Index (an index typically used to compare species composition across sites), treating variables as species and individual analyses as sites. To generate an individual Sorensen's value for each analysis required calculating the pairwise Sorensen's value for all pairs of analyses (of the same dataset), and then taking the average across these Sorensen's values for each analysis. We calculated the Sorensen's index values using the *betapart* package [10] in R:

$$\beta Sorensen = \frac{b + c}{2a + b + c}$$

where a is the number of variables common to both analyses, b is the number of variables that occur in the first analysis but not in the second and c is the number of variables that occur in the second analysis. We then used the per-model average Sorensen's index value as an independent variable to predict the deviation score in a general linear model, and included no random effect since each analysis is included only once, in R [87]:

Additional explanation:

When we planned this analysis, we anticipated that analysts would identify a single primary effect from each model, so that each model would appear in the analysis only once. Our expectation was incorrect because some analysts identified >1 effect per analysis, but we still chose to specify our model as registered and not use a random effect. This is because most models produced only one effect and so we expected that specifying a random effect to account for the few cases where >1 effect was included for a given model would prevent model convergence.

Note that this analysis contrasts with the analyses in which we used reviewer ratings as predictors because in the analyses with reviewer ratings, each effect appeared in the analysis approximately four times due to multiple reviews of each analysis, and so it was much more important to account for that variance through a random effect.

$$DeviationScore_j \sim \beta Sorensen_j$$

Next, we assessed the relationship between the inclusion of random effects in the analysis and the deviation from the mean effect size. We anticipated that most analysts would use random effects in a mixed model framework, but if we were wrong, we wanted to evaluate the differences in outcomes when using random effects versus not using random effects. Thus, if there were at least 5 analyses that did and 5 analyses that did not include random effects, we would add a binary predictor variable "random effects included (yes/no)" to our set of univariate analyses and would add this predictor variable to our multivariate model described below. This standard was only met for the *Eucalyptus* analyses, and so we only examined inclusion of random effects as a predictor variable in meta-analysis of this set to analyses.

Finally, we conducted a multivariate analysis with the five predictors described above (peer ratings 0–100 and peer ratings of publishability 1–4; both original and revised and Sorensen's index, plus a sixth for *Eucalyptus*, presence / absence of random effects) with random effects of analysis identity and reviewer identity in the *lme4* package in R [11, 87]. We had stated here in the text that we would use only the revised (final) peer ratings in this analysis, so the absence of the initial ratings is not a deviation from our plan:

$$\text{DeviationScore}_j = BoxCox\left(\text{DeviationFromMean}_j\right)$$

$$\text{DeviationScore}_{ij} \sim \text{RatingContinuous}_{ij} + \text{RatingCategorical}_{ij} + \beta\text{Sorensen}_j + \text{ReviewerID}_i + \text{Effect ID}_j$$

$$\text{ReviewerID}_i \sim N(0, \sigma_i^2)$$

$$\text{EffectID}_j \sim N(0, \sigma_j^2)$$

We conducted all the analyses described above eight times; for each of the four responses (Z_r, y_{25}, y_{50}, y_{75}) one time for each of the two datasets.

We have publicly archived all relevant data, code, and materials on the Open Science Framework (https://osf.io/mn5aj/). Archived data includes the original datasets distributed to all analysts, any edited versions of the data analyzed by individual groups, and the data we analyzed with our meta-analyses, which include the effect sizes derived from separate analyses, the statistics describing variation in model structure among analyst groups, and the anonymized answers to our surveys of analysts and peer reviewers. Similarly, we have archived both the analysis code used for each individual analysis (where available) and the code from our meta-analyses. We have also archived copies of our survey instruments from analysts and peer reviewers.

Our rules for excluding data from our study were as follows. We excluded from our synthesis any individual analysis submitted after we had completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We also excluded any individual analysis that did not produce an outcome that could be interpreted as an answer to our primary question (as posed above) for the respective dataset. For instance, this means that in the case of the data on blue tit chick growth, we excluded any analysis that did not include something that can be interpreted as growth or size as a dependent (response) variable, and in the case of the *Eucalyptus* establishment data, we excluded any analysis that did not include a measure of grass cover among the independent (predictor) variables. Also, as described above, any analysis that could not produce an effect that could be converted to a signed Z_r was excluded from analyses of Z_r.

Preregistration Deviation:

Some analysts had difficulty implementing our instructions to derive the out-of-sample predictions, and in some cases (especially for the *Eucalyptus* data), they submitted predictions with implausibly extreme values. We believed these values were incorrect and thus made the conservative decision to exclude out-of-sample predictions where the estimates were > 3 standard deviations from the mean value from the full dataset provided to teams for analysis.

Additional explanation: Best practices in many-analysts research

After we initiated our project, a paper was published outlining best practices in many-analysts studies [1]. Although we did not have access to this document when we implemented our project, our study complies with these practices nearly completely. The one exception is that although we requested analysis code from analysts, we did not require submission of code.

Additional explanation: unregistered analyses

1. Evaluating model fit.

We evaluated all fitted models using the performance::performance() function from the *performance* package [60] and the glance() function from the *broom.mixed* package [13]. For all models, we calculated the square root of the residual variance (Sigma) and the root mean squared error (RMSE). For GLMMs performance::performance() calculates the marginal and conditional R^2values as well as the contribution of random effects (ICC), based on [76]. The conditional R^2 accounts for both the fixed and random effects, while the marginal R^2 considers only the variance of the fixed effects. The contribution of random effects is obtained by subtracting the marginal R^2 from the conditional R^2.

2. Exploring outliers and analysis quality.

After seeing the forest plots of *Zr* values and noticing the existence of a small number of extreme outliers, especially from the *Eucalyptus* analyses, we wanted to understand the degree to which our heterogeneity estimates were influenced by these outliers. To explore this question, we removed the highest two and lowest two values of *Zr* in each dataset and re-calculated our heterogeneity estimates.

To help understand the possible role of the quality of analyses in driving the heterogeneity we observed among estimates of *Zr*, we created forest plots and recalculated our heterogeneity estimates after removing all effects from analysis teams that had received at least one rating of "deeply flawed and unpublishable" and then again after removing all effects from analysis teams with at least one rating of either "deeply flawed and unpublishable" or "publishable with major revisions". We also used self-identified levels of statistical expertise to examine heterogeneity when we retained analyses only from analysis teams that contained at least one member who rated themselves as "highly proficient" or "expert" (rather than "novice" or "moderately proficient") in conducting statistical analyses in their research area in our intake survey.

Additionally, to assess potential impacts of highly collinear predictor variables on estimates of *Zr* in blue tit analyses, we created forest plots (Supplementary Material B, Figure B.5) and recalculated our heterogeneity estimates after we removed analyses that contained the brood count after manipulation and the highly correlated (correlation of 0.89, Supplementary Material D, Figure D.2) brood count at day 14. This removal included the one effect based on a model that contained both these variables and a third highly correlated variable, the estimate of number of chicks fledged (the only model that included the estimate of number of chicks fledged). We did not conduct a similar analysis for the *Eucalyptus* dataset because there were no variables highly collinear with the primary predictors (grass cover variables) in that dataset (Supplementary Material D, Figure D.1).

3. Exploring possible impacts of lower quality estimates of degrees of freedom.

Our meta-analyses of variation in Z_r required variance estimates derived from estimates of the degrees of freedom in original analyses from which Z_r estimates were derived. While processing the estimates of degrees of freedom submitted by analysts, we identified a subset of these estimates in which we had lower confidence because two or more effects from the same analysis were submitted with identical degrees of freedom. We therefore conducted a second set of (more conservative) meta-analyses that excluded these Z_r estimates with identical estimates of degrees of freedom and we present these analyses in the supplement.

Step 6: Facilitated discussion and collaborative write-up of manuscript

We planned for analysts and initiating authors to discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

Preregistration deviation:

As described above, due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We built an R package, ManyEcoEvo:: to conduct the analyses described in this study [38], which can be downloaded from https://github.com/egouldo/ManyEcoEvo/ to reproduce our analyses or replicate the analyses described here using alternate datasets. Data cleaning and preparation of analysis data, as well as the analysis, is conducted in R [87] reproducibly using the targets package ([57]). This data and analysis pipeline is stored in the ManyEcoEvo:: package repository and its outputs are made available to users of the package when the library is loaded.

The full manuscript, including further analysis and presentation of results is written in Quarto [2]. The source code to reproduce the manuscript is hosted at https://github.com/egouldo/ManyAnalysts/ [39], and the rendered version of the source code may be viewed at https://egouldo.github.io/ManyAnalysts/. All R packages and their versions used in the production of the manuscript are listed in Table 7 at the end of this paper.

Results

Summary statistics

In total, 173 analyst teams, comprising 246 analysts, contributed 182 usable analyses (compatible with our meta-analyses and provided with all information needed for inclusion) of the two datasets examined in this study which yielded 215 effects. Analysts produced 134 distinct effects that met our criteria for inclusion in at least one of our meta-analyses for the blue tit dataset. Analysts produced 81 distinct effects meeting our criteria for inclusion for the *Eucalyptus* dataset. Excluded analyses and effects either did not answer our specified biological questions, were submitted with insufficient information for inclusion in our meta-analyses, or were incompatible with production of our effect size(s). We expected cases of this final scenario (incompatible analyses), for instance we cannot extract a Z_r from random forest models, which is why we analyzed two distinct types of effects, Z_r and out-of-sample predictions. Some effects only provided sufficient information for a subset of analyses and were only included in that subset. For both datasets, most submitted analyses incorporated mixed effects. Submitted analyses of the blue tit dataset typically specified normal error and analyses of the *Eucalyptus* dataset typically specified a non-normal error distribution (Supplementary Material A, Table A.1).

For both datasets, the composition of models varied substantially in regards to the number of fixed and random effects, interaction terms, and the number of data points used, and these patterns differed somewhat between the blue tit and *Eucalyptus* analyses (See Supplementary Material A Table A. 2). Focusing on the models included in the Z_r analyses (because this is the larger sample), blue tit models included a similar number of fixed effects on average (mean 5.2 ± 2.92 SD, range 1 to 19) as *Eucalyptus* models (mean 5.01 ± 3.83 SD, range 1 to 13), but the standard deviation in the number of fixed effects was somewhat larger in the *Eucalyptus* models. The average number of interaction terms was much larger for the blue tit models (mean 0.44 ± 1.11 SD, range 0 to 10) than for the *Eucalyptus* models (mean 0.16 ± 0.65 SD, range 0 to 5), but still under 0.5 for both, indicating that most models did not contain interaction terms. Blue tit models also contained more random effects (mean 3.53 ± 2.08 SD, range 0 to 10) than *Eucalyptus* models (mean 1.41 ± 1.09 SD, range 0 to 4). The maximum possible sample size in the blue tit dataset (3720 nestlings) was an order of magnitude larger than the maximum possible in the *Eucalyptus* dataset (351 plots), and the means and standard deviations of the sample size used to derive the effects eligible for our study were also an order of magnitude greater for the blue tit dataset (mean 2611.09 ± 937.48 SD, range 76 to 76) relative to the *Eucalyptus* models (mean 298.43 ± 106.25 SD, range 18 to 351). However, the standard deviation in sample size from the *Eucalyptus* models was heavily influenced by a few cases of dramatic sub-setting (described below). Approximately three quarters of *Eucalyptus* models used sample sizes within 3% of the maximum. In contrast, fewer than 20% of blue tit models relied on sample sizes within 3% of the maximum, and approximately 50% of blue tit models relied on sample sizes 29% or more below the maximum.

Analysts provided qualitative descriptions of the conclusions of their analyses. Each analysis team provided one conclusion per dataset. These conclusions could take into account the results of any formal analyses completed by the team as well as exploratory and visual analyses of the data. Here we summarize all qualitative responses, regardless of whether we had sufficient information to use the corresponding model results in our quantitative analyses below. We classified these conclusions into the categories summarized below (Table 1):

- *Mixed*: some evidence supporting a positive effect, some evidence supporting a negative effect
- *Conclusive negative*: negative relationship described without caveat
- *Qualified negative*: negative relationship but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive none*: analysts interpret the results as conclusive of no effect
- *Qualified none*: analysts describe finding no evidence of a relationship but they describe the potential for an undetected effect
- *Qualified positive*: positive relationship described but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive positive*: positive relationship described without caveat

For the blue tit dataset, most analysts concluded that there was negative relationship between measures of sibling competition and nestling growth, though half the teams expressed qualifications or described effects as mixed or absent. No analysts concluded that there was a positive relationship even though some individual effect sizes were positive, apparently because all analysts who produced effects indicating positive relationships also produced effects indicating negative relationships and therefore described their results as qualified, mixed, or absent. For the *Eucalyptus* dataset, there was a broader spread of conclusions with at least one analyst team providing conclusions consistent with each conclusion category. The most common conclusion for the *Eucalyptus* dataset was that there was no relationship between grass cover and *Eucalyptus* recruitment (either conclusive or qualified description of no relationship), but more than half the teams concluded that there were effects; negative, positive, or mixed.

Distribution of effects

Effect sizes (Zr)

Although the majority (118 of 131) of the usable *Zr* effects from the blue tit dataset found nestling growth decreased with sibling competition, and the meta-analytic mean $\overline{Z}_r$ (Fisher's transformation of the correlation coefficient) was convincingly negative (−0.35 ± 0.06 95%CI), there was substantial variability in the strength and the direction of this effect. Z_r ranged from −1.55 to 0.38, and approximately continuously from −0.93 to 0.19 (Fig. 2a and Table 4), and of the 118 effects with negative slopes, 93 had confidence intervals excluding 0. Of the 13 with positive slopes indicating increased nestling growth in the presence of more siblings, 2 had confidence intervals excluding zero (Fig. 2a).

Meta-analysis of the *Eucalyptus* dataset also showed substantial variability in the strength of effects as measured by Z_r, and unlike with the blue tits, a notable lack of consistency in the direction of effects (Fig. 2b, Table 4). Z_r ranged from −4.47 (Supplementary Material A, Figure A.2), indicating a strong tendency for reduced *Eucalyptus* seedling success as grass cover increased, to 0.39, indicating the opposite. Although the range of reported effects skewed strongly negative, this was due to a small number of substantial outliers. Most values of Z_r were relatively small with values <|0.2| and the meta-analytic mean effect size was close to zero (−0.09 ± 0.12 95%CI). Of the 79 effects, fifty-three had confidence intervals overlapping zero, approximately a quarter (fifteen) crossed the traditional threshold of statistical significance indicating a negative relationship between grass cover and seedling success, and eleven crossed the significance threshold indicating a positive relationship between grass cover and seedling success (Fig. 2b).

Out-of-sample predictions (yi)

As with the effect size Z_r, we observed substantial variability in the size of out-of-sample predictions derived from the analysts' models. Blue tit predictions (Fig. 3a), which were *z*-score-standardized to accommodate the use of different response variables, always ranged far in excess of one standard deviation. In the y_{25} scenario, model predictions ranged from −1.84 to 0.42 (a range of 2.68 standard deviations), in the y_{50} they ranged from −0.52 to 1.08 (a range of 1.63 standard deviations), and in the y_{75} scenario they ranged from −0.03 to 1.59 (a range of 1.9 standard deviations). As should be expected given the existence of both negative and positive Z_r values, all three out-of-sample scenarios produced both negative and positive predictions, although as with the Z_r values, there is a clear trend for scenarios with more siblings to be associated with smaller nestlings. This is supported by the meta-analytic means of these three sets of predictions which were −0.66 (95%CI −0.82–0.5) for the y_{25}, 0.34 (95%CI 0.2–0.48) for the y_{50}, and 0.67 (95%CI 0.57–0.77) for the y_{75}.

Eucalyptus out-of-sample predictions also varied substantially (Fig. 3b), but because they were not *z*-score-standardized and are instead on the original count scale, the types of interpretations we can make differ. The predicted *Eucalyptus* seedling counts per 15 × 15 m plot for the y_{25} scenario ranged from 0.04 to 26.99, for the y_{50} scenario ranged from 0.04 to 44.34, and for the y_{75} scenario they ranged from 0.03 to 61.34. The meta-analytic mean predictions for these three scenarios were similar; 1.27 (95%CI 0.59–2.3) for the y_{25}, 2.92 (95%CI 0.98–3.89) for the y_{50}, and 2.92 (95%CI 1.59–4.9) for the y_{75} scenarios respectively.

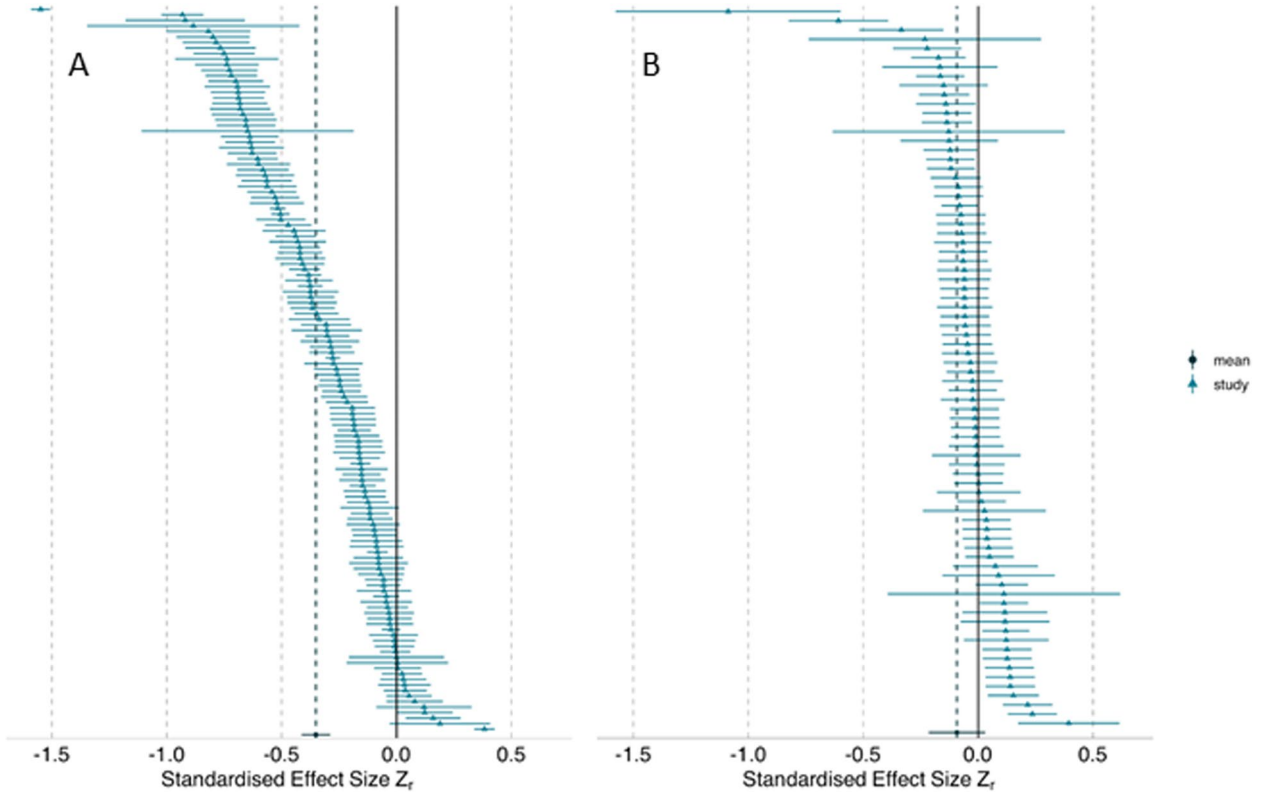


Fig. 2 Forest plots of meta-analytic estimated standardized effect sizes (Z_r, blue triangles) and their 95% confidence intervals for each effect size included in the meta-analysis model. **A** Blue tit analyses: Points where Z_r are less than 0 indicate analyses that found a negative relationship between sibling number and nestling growth. **B** *Eucalyptus* analyses: Points where Z_r are less than 0 indicate a negative relationship between grass cover and Eucalyptus seedling success. The meta-analytic mean effect size is denoted by a black circle and a dashed vertical line, with error bars also representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0, indicating no relationship between the test variable and the response variable. Note that the *Eucalyptus* plot omits one extreme outlier with the value of −4.47 (Supplementary Material A, Figure A.2) in order to standardize the *x*-axes on these two panels

Table 1 Tallies of analysts' qualitative answers to the research questions addressed by their analyses

Dataset	Mixed	Negative Conclusive	Negative Qualified	None Conclusive	None Qualified	Positive Qualified	Positive Conclusive
blue tit	5	37	27	4	1	0	0
Eucalyptus	8	6	12	19	12	4	2

Quantifying heterogeneity

Effect sizes (Zr)

We quantified both absolute (τ^2) and relative (I^2) heterogeneity resulting from analytical variation. Both measures suggest that substantial variability among effect sizes was attributable to the analytical decisions of analysts.

The total absolute level of variance beyond what would typically be expected due to sampling error, τ^2 (Table 2), among all usable blue tit effects was 0.08 and for *Eucalyptus* effects was 0.27. This is similar to or exceeding the median value (0.105) of τ^2 found across 31 recent meta-analyses (calculated from the data in [121]). The similarity of our observed values to values from meta-analyses of different studies based on different data suggests the potential for a large portion of heterogeneity to arise from analytical decisions. For further discussion of interpretation of τ^2 in our study, please consult discussion of post hoc analyses below.

In our analyses, I^2 is a plausible index of how much more variability among effect sizes we have observed, as a proportion, than we would have observed if sampling error were driving variability. We discuss our

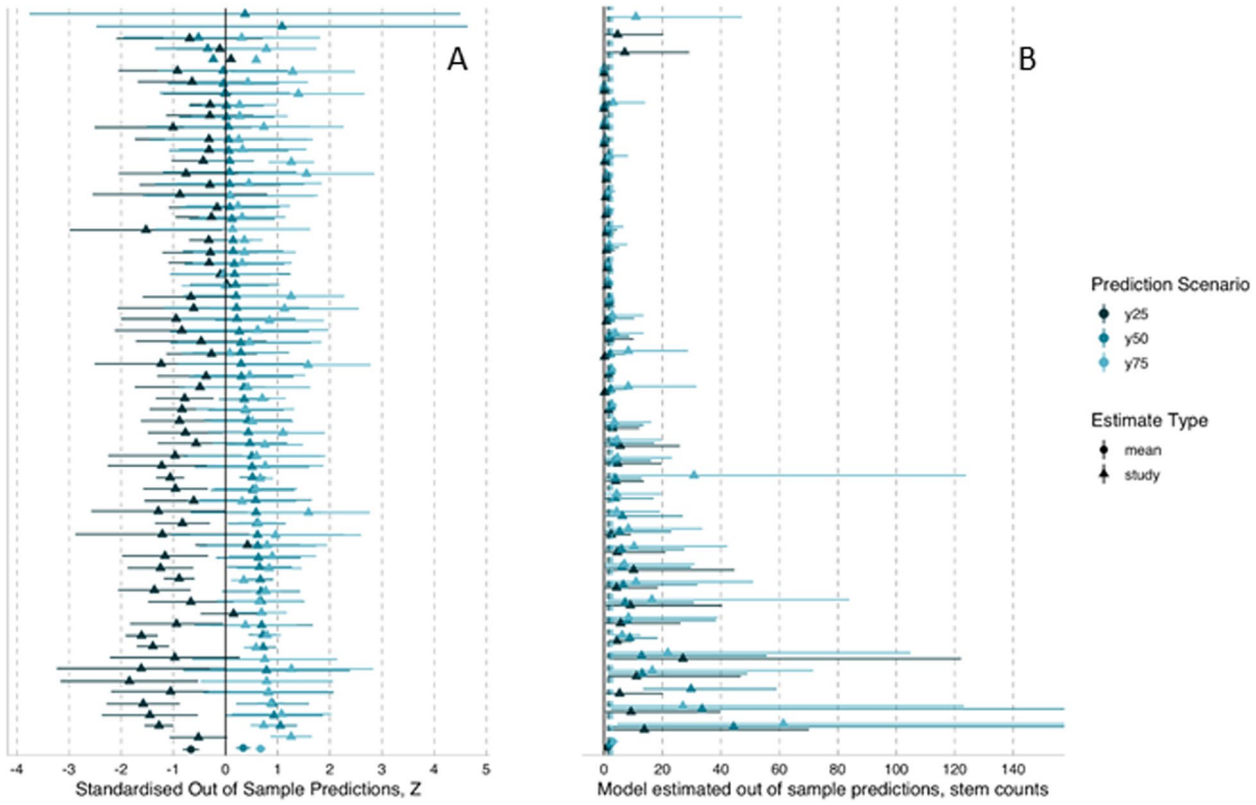


Fig. 3 Forest plot of meta-analytic estimated out-of-sample predictions. **A** Standardized (*z*-score) blue tit out-of-sample predictions, y_i. **B** Response-scale (stem counts) *Eucalyptus* out-of-sample predictions. Triangles represent individual estimates. Circles represent the meta-analytic mean for each prediction scenario. Dark-blue points correspond to y_{25} scenario, medium-blue points correspond to the y_{50} scenario, while light blue points correspond to the y_{75} scenario. Error bars are 95% confidence intervals. Note that, for the *Eucalyptus* analysis, outliers (observations more than 3 SD above the mean) have been removed prior to model fitting and do not appear on this figure. The *x*-axis is truncated to approximately 140, and thus some error bars are incomplete. See Supplementary Material B, Figure B.6 for full figure

interpretation of I^2 further in the methods, but in short, it is a useful metric for comparison to values from published meta-analyses and provides a plausible value for how much heterogeneity could arise in a normal meta-analysis with similar sample sizes due to analytical variability alone. In our study, total I^2 for the blue tit Z_r estimates was extremely large, at 97.61%, as was the *Eucalyptus* estimate (98.59% Table 2).

Although the overall I^2 values were similar for both *Eucalyptus* and blue tit analyses, the relative composition of that heterogeneity differed. For both datasets, the majority of heterogeneity in Z_r was driven by differences among effects as opposed to differences among teams, though this was more prominent for the *Eucalyptus* dataset, where nearly all of the total heterogeneity was driven by differences among effects (91.7%) as opposed to differences among teams (6.89%) (Table 2).

Out-of-sample predictions (yi)

We observed substantial heterogeneity among out-of-sample estimates, but the pattern differed somewhat from the Z_r values (Table 3). Among the blue tit predictions, I^2 ranged from medium-high for the y_{25} scenario (68.54) to low (27.9) for the y_{75} scenario. Among the *Eucalyptus* predictions, I^2 values were uniformly high (>82%). For both datasets, most of the existing heterogeneity among predicted values was attributable to among-team differences, with the exception of the y_{50} analysis of the *Eucalyptus* dataset. We are limited in our interpretation of τ^2 for these estimates because, unlike for the Z_r estimates, we have no benchmark for comparison with other meta-analyses.

Post hoc analysis: Exploring outlier characteristics and the effect of outlier removal on heterogeneity

Effect sizes (Zr)

The outlier *Eucalyptus Zr* values were striking and merited special examination. The three negative outliers had very low sample sizes that were based on either small subsets of the dataset or, in one case, extreme aggregation of data. The outliers associated with small subsets had sample sizes (n= 117, 90, 18) that were less than

Table 2 Heterogeneity in the estimated effects Z_r for meta-analyses of: the full dataset, as well as from post hoc analyses wherein analyses with outliers are removed, analyses with effects from analysis teams with at least one "unpublishable" rating are excluded, analyses receiving at least one "major revisions" rating or worse excluded, analyses from teams with at least one analyst self-rated as "highly proficient" or "expert" in statistical analysis are included, and (blue tit only) analyses that did not included the pair of highly collinear predictors together. τ^2_{Team} is the absolute heterogeneity for the random effect Team. $\tau^2_{\text{Effect ID}}$ is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{\text{Team, EffectID}}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences

Dataset	N_{Obs}	τ^2_{Total}	τ^2_{Team}	τ^2_{EffectID}	I^2_{Total}	I^2_{Team}	$I^2_{\text{Team, EffectID}}$
All Analyses							
Eucalyptus	79	0.27	0.02	0.25	98.59%	6.89%	91.70%
blue tit	131	0.08	0.03	0.05	97.61%	36.71%	60.90%
Blue tit analyses containing highly collinear predictors removed							
blue tit	117	0.07	0.04	0.03	96.92%	58.18%	38.75%
All analyses, outliers removed							
Eucalyptus	75	0.01	0.00	0.01	66.19%	19.25%	46.94%
blue tit	127	0.07	0.04	0.02	96.84%	64.63%	32.21%
Analyses receiving at least one "Unpublishable" rating removed							
Eucalyptus	55	0.01	0.01	0.01	79.74%	28.31%	51.43%
blue tit	109	0.08	0.03	0.05	97.52%	35.68%	61.84%
Analyses receiving at least one "Unpublishable" and or "Major Revisions" rating removed							
Eucalyptus	13	0.03	0.03	0.00	88.91%	88.91%	0.00%
blue tit	32	0.14	0.01	0.13	98.72%	5.17%	93.55%
Analyses from teams with highly proficient or expert data analysts							
Eucalyptus	34	0.58	0.02	0.56	99.41%	3.47%	95.94%
blue tit	89	0.09	0.03	0.06	97.91%	31.43%	66.49%

half of the total possible sample size of 351. The case of extreme aggregation involved averaging all values within each of the 351 sites in the dataset.

Surprisingly, both the largest and smallest effect sizes in the blue tit analyses (Fig. 2a) come from the same analyst (anonymous ID: "Adelong"), with identical models in terms of the explanatory variable structure, but with different response variables. However, the radical change in effect was primarily due to collinearity with covariates. The primary predictor variable (brood count after manipulation) was accompanied by several collinear variables, including the highly collinear (correlation of 0.89 Supplementary Material D, Figure D.2) covariate (brood count at day 14) in both analyses. In the analysis of nestling weight, brood count after manipulation showed a strong positive partial correlation with weight after controlling for brood count at day 14 and treatment category (increased, decreased, unmanipulated). In that same analysis, the most collinear covariate (the day 14 count) had a negative partial correlation with weight. In the analysis with tarsus length as the response variable, these partial correlations were almost identical in absolute magnitude, but reversed in sign and so brood count after manipulation was now the collinear predictor with the negative relationship. The two models were therefore very similar, but the two collinear predictors simply switched roles, presumably because a subtle difference in the distribution of weight and tarsus length data.

When we dropped the *Eucalyptus* outliers, I^2 decreased from high (98.59 %), using [43] suggested benchmark, to between moderate and high (66.19 %, Table 2). However, more notably, τ^2 dropped from 0.27 to 0.01, indicating that, once outliers were excluded, the observed variation in effects was similar to what we would expect if sampling error were driving the differences among effects (since τ^2 is the variance beyond that driven by sampling error). The interpretation of this value of τ^2 in the context of our many-analyst study is somewhat different than a typical meta-analysis, however, since in our study (especially for *Eucalyptus*, where most analyses used almost exactly the same data points), there is almost no role for sampling error in driving the observed differences among

Table 3 Heterogeneity among the out-of-sample predictions y_i for both blue tit and *Eucalyptus* datasets. τ^2_{Team} is the absolute heterogeneity for the random effect Team. T^2_{EffectID} is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{\text{Team,Effect ID}}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences

Prediction Scenario	N_{Obs}	T_{Total}	T^2_{Team}	T^2_{EffectID}	I^2_{Total}	I^2_{Team}	$I^2_{\text{Team,EffectID}}$
blue tit							
y25	63	0.23	0.11	0.03	68.54%	53.43%	15.11%
y50	60	0.23	0.06	0.00	50%	46.29%	3.71%
y75	63	0.23	0.02	0.00	27.9%	27.89%	0.01%
Eucalyptus							
y25	38	5.75	1.48	0.68	86.93%	59.54%	27.39%
y50	38	5.75	1.32	0.83	89.63%	55%	34.64%
y75	38	5.75	1.03	0.41	80.19%	57.41%	22.78%

Table 4 Estimated mean value of the standardized correlation coefficient, $\overline{Z}_r$, along with its standard error and 95% confidence intervals. We re-computed the meta-analysis for different post hoc subsets of the data: All eligible effects, removal of effects from blue tit analyses that contained a pair of highly collinear predictor variables, removal of effects from analysis teams that received at least one peer rating of "deeply flawed and unpublishable", removal of any effects from analysis teams that received at least one peer rating of either "deeply flawed and unpublishable" or "publishable with major revisions", inclusion of only effects from analysis teams that included at least one member who rated themselves as "highly proficient" or "expert" at conducting statistical analyses in their research area

Dataset	$\widehat{\mu}$	$SE[\widehat{\mu}]$	95% CI	Statistic	*p*
All analyses					
Eucalyptus	−0.09	0.06	[−0.22,0.03]	−1.47	0.14
blue tit	−0.35	0.03	[−0.41,−0.29]	−11.02	<0.001
Blue tit analyses containing highly collinear predictors removed					
blue tit	−0.36	0.03	[−0.42,−0.29]	−10.97	<0.001
All analyses, outliers removed					
Eucalyptus	−0.03	0.01	[−0.06,0.00]	−2.23	0.026
blue tit	−0.36	0.03	[−0.42,−0.30]	−11.48	<0.001
Analyses receiving at least one "Unpublishable" rating removed					
Eucalyptus	−0.02	0.02	[−0.07,0.02]	−1.15	0.3
blue tit	−0.36	0.03	[−0.43,−0.30]	−10.82	<0.001
Analyses receiving at least one "Unpublishable" and or "Major Revisions" rating removed					
Eucalyptus	−0.04	0.05	[−0.15,0.07]	−0.77	0.4
blue tit	−0.37	0.07	[−0.51,−0.23]	−5.34	<0.001
Analyses from teams with highly proficient or expert data analysts					
Eucalyptus	−0.17	0.13	[−0.43,0.10]	−1.24	0.2
blue tit	−0.36	0.04	[−0.44,−0.28]	−8.93	<0.001

the estimates. Thus, rather than concluding that the variability we observed among estimates (after removing outliers) was due only to sampling error (because τ^2 became small: 10% of the median from [121], we instead conclude that the observed variability, which must be due to the divergent choices of analysts rather than sampling error, is approximately of the same magnitude as what we would have expected if, instead, sampling error, and not analytical heterogeneity, were at work. Conversely, dropping outliers from the set of blue tit effects did not meaningfully reduce I^2, and only modestly reduced τ^2 (Table 2). Thus, effects at the extremes of the distribution were much stronger contributors to total heterogeneity for effects from analyses of the *Eucalyptus* than for the blue tit dataset.

Out-of-sample predictions (yi)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller and the pattern of outliers differed.

Post hoc analysis: Exploring the effect of removing analyses with poor peer ratings on heterogeneity

Effect sizes (Zr)

Removing poorly rated analyses had limited impact on the meta-analytic means (Supplementary Material B, Figure B.3). For the *Eucalyptus* dataset, the meta-analytic mean shifted from −0.09 to −0.02 when effects from analyses rated as unpublishable were removed, and to −0.04 when effects from analyses rated, at least once, as unpublishable or requiring major revisions were removed. Further, the confidence intervals for all of these means overlapped each of the other means (Table 4). We

saw similar patterns for the blue tit dataset, with only small shifts in the meta-analytic mean, and confidence intervals of all three means overlapping each other mean (Table 4). Refitting the meta-analysis with a fixed effect for categorical ratings also showed no indication of differences in group meta-analytic means due to peer ratings (Supplementary Material B, Figure B.1).

For the blue tit dataset, removing poorly rated analyses led to only negligible changes in I^2_{Total} and relatively minor impacts on τ^2. However, for the *Eucalyptus* dataset, removing poorly rated analyses led to notable reductions in I^2_{Total} and substantial reductions in τ^2. When including all analyses, the *Eucalyptus* I^2_{Total} was 98.59% and τ^2 was 0.27, but eliminating analyses with ratings of "unpublishable" reduced I^2_{Total} to 79.74% and τ^2 to 0.01, and removing also those analyses "needing major revisions" left I^2_{Total} at 88.91% and τ^2 at 0.03 (Table 2). Additionally, the allocations of I^2 to the team versus individual effect were altered for both blue tit and *Eucalyptus* meta-analyses by removing poorly rated analyses, but in different ways. For blue tit meta-analysis, between a third and two-thirds of the total I^2 was attributable to among-team variance in most analyses until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost all remaining heterogeneity was attributable to among-effect differences. In contrast, for *Eucalyptus* meta-analysis, the among-team component of I^2 was less than third until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost 90% of heterogeneity was attributable to differences among teams.

Out-of-sample predictions (yi)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller and our ability to interpret heterogeneity values for these analyses was limited

Post hoc analysis: exploring the effect of including only analyses conducted by analysis teams with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area

Effect sizes (Zr)

Including only analyses conducted by teams that contained at least one member who rated themselves as "highly proficient" or "expert" in conducting the relevant statistical methods had negligible impacts on the meta-analytic means (Table 4), the distribution of Z_r effects (Supplementary Material B, Figure B.4), or heterogeneity estimates (Table 2), which remained extremely high.

Out-of-sample predictions (yi)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller.

Post hoc analysis: exploring the effect of excluding estimates of Z_r in which we had reduced confidence

As described in our addendum to the methods, we identified a subset of estimates of Z_r in which we had less confidence because of features of the submitted degrees of freedom. Excluding these effects in which we had lower confidence had minimal impact on the meta-analytic mean and the estimates of total I^2 and τ^2 for both blue tit and *Eucalyptus* meta-analyses, regardless of whether outliers were also excluded (Supplementary Material B, Table B. 1).

Post hoc analysis: exploring the effect of excluding effects from blue tit models that contained two highly collinear predictors

Effect sizes (Zr)

Excluding effects from blue tit models that contained the two highly collinear predictors (brood count after manipulation and brood count at day 14) had negligible impacts on the meta-analytic means (Table 4), the distribution of Z_r effects (Supplementary Material B, Figure B.5), or heterogeneity estimates (Table 2), which remained high.

Out-of-sample predictions

Inclusion of collinear predictors does not harm model prediction, and so we did not conduct these post hoc analyses.

Explaining variation in deviation scores

None of the pre-registered predictors explained substantial variation in deviation among submitted statistical effects from the meta-analytic mean (Tables 5 and 6).

Deviation scores as explained by reviewer ratings

Effect sizes (Zr)

We obtained reviews from 153 reviewers who reviewed analyses for a mean of 3.27 (range 1–11) analysis teams. Analyses of the blue tit dataset received a total of 240 reviews, each was reviewed by a mean of 3.87 (SD 0.71, range 3–5) reviewers. Analyses of the *Eucalyptus* dataset received a total of 178 reviews, each was reviewed by a mean of 4.24 (SD 0.79, range 3–6) reviewers. We tested for inter-rater-reliability (IRR) to examine how similarly reviewers reviewed each analysis and found approximately no agreement among reviewers. When

considering continuous ratings, IRR was 0.01, and for categorical ratings, IRR was −0.14.

Many of the models of deviation as a function of peer ratings faced issues of failure to converge or singularity due to sparse design matrices with our pre-registered random effects (Effect ID and Reviewer ID) (see Supplementary Material C). These issues persisted after increasing the tolerance and changing the optimizer. For both *Eucalyptus* and blue tit datasets, models with continuous ratings as a predictor were singular when both pre-registered random effects were included.

When using both categorical and continuous ratings as predictors, only models converged and allowed 95% confidence intervals to be calculated when specifying Reviewer ID as a random effect. The categorical ratings model had a R^2_C of 0.09 and a R^2_M of 0.01, the continuous ratings model had a R^2_C of 0.09 and a R^2_M of 0.01 for the blue tit dataset and a R^2_C of 0.12 and a R^2_M of 0.01 for the *Eucalyptus* dataset. Neither continuous nor categorical reviewer ratings of the analyses meaningfully predicted deviance from the meta-analytic mean (Table 6, Fig. 4). We re-ran the multi-level meta-analysis with a fixed effect for the categorical publishability ratings and found no difference in mean standardized effect sizes among publishability ratings (Supplementary Material B, Figure B.1).

Out-of-sample predictions (yi)

Some models of the influence of reviewer ratings on out-of-sample predictions (y_i) had issues with convergence and singularity of fit (see Supplementary Material C, Table C.3) and those models that converged and were not singular showed no strong relationship (Supplementary Material C, Figure C.2, Supplementary Material C, Figure C.3), as with the Z_r analyses.

Deviation scores as explained by the distinctiveness of variables in each analysis

Effect sizes (Zr)

We employed Sorensen's index to calculate the distinctiveness of the set of predictor variables used in each model (Fig. 5). The mean Sorensen's score for blue tit analyses was 0.59 (SD 0.1, range 0.43–0.86), and for *Eucalyptus* analyses was 0.69 (SD 0.08, range 0.55–0.98).

We found no meaningful relationship between distinctiveness of variables selected and deviation from the meta-analytic mean (Table 6, Fig. 5) for either blue tit (mean 0.42, 95%CI −0.49,1.34) or *Eucalyptus* effects (mean 0.18, 95%CI −2.78,3.14).

Out-of-sample predictions (yi)

As with the Z_r estimates, we did not observe any convincing relationships between deviation scores of

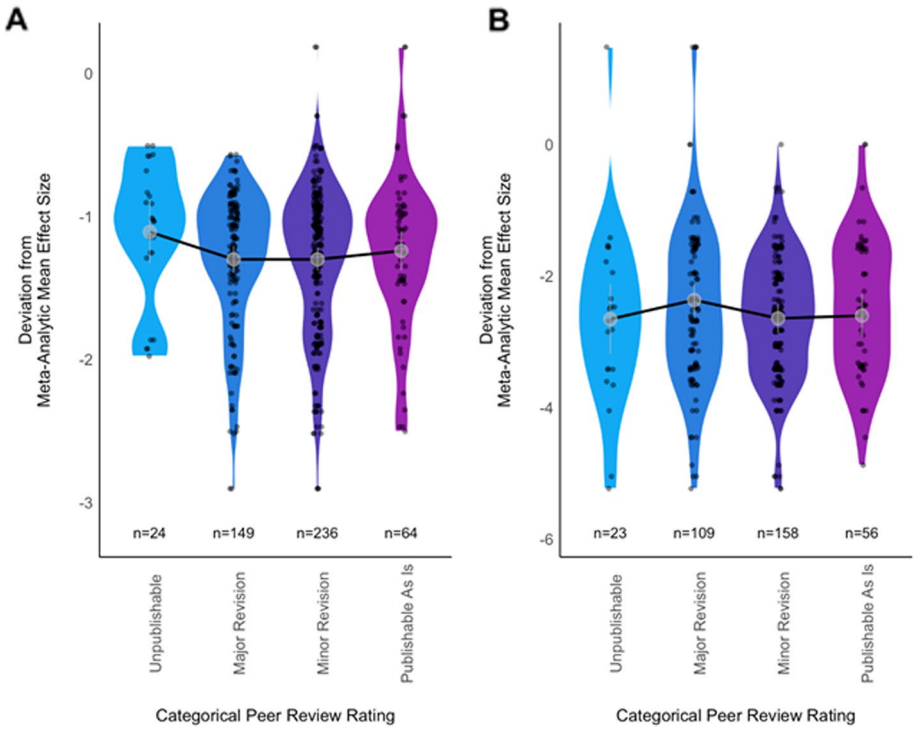


Fig. 4 Violin plot of Box-Cox transformed deviation from meta-analytic mean $\overline{Z}_r$ as a function of categorical peer rating. Grey points for each rating group denote model-estimated marginal mean deviation, and error bars denote 95%CI of the estimate. **A** Blue tit dataset. **B***Eucalyptus* dataset

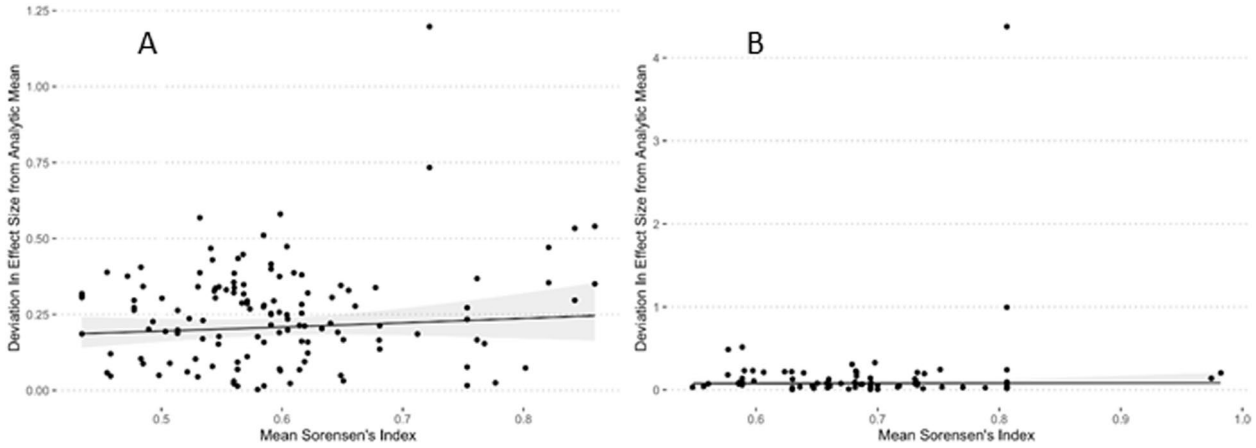


Fig. 5 Fitted model of the Box-Cox-transformed deviation score (deviation in effect size from meta-analytic mean) as a function of the mean Sorensen's index showing distinctiveness of the set of predictor variables. Grey ribbons on predicted values are 95% CI's. **A** blue tit dataset. **B***Eucalyptus* dataset

out-of-sample predictions and Sorensen's index values (see Supplementary Material C4.1).

Deviation scores as explained by the inclusion of random effects

Effect sizes (Zr)

There were only three blue tit analyses that did not include random effects, which is below the pre-registered threshold for fitting a model of the Box-Cox transformed deviation from the meta-analytic mean as a function of whether the analysis included random-effects. However, 17 *Eucalyptus* analyses included only fixed effects, which crossed our pre-registered threshold. Consequently, we performed this analysis for the *Eucalyptus* dataset only. There was no relationship between random-effect inclusion and deviation from meta-analytic mean among the *Eucalyptus* analyses (Tables 6, Fig. 6).

Out-of-sample predictions (yi)

As with the Z_r estimates, we did not examine the possibility of a relationship between the inclusion of random effects and the deviation scores of the blue tit out-of-sample predictions. When we examined the possibility of this relationship for the *Eucalyptus* effects, we found consistent evidence of somewhat higher Box-Cox-transformed deviation values for models including a random effect, meaning the models including random effects averaged slightly higher deviation from the meta-analytic means (Supplementary Material C, Figure C.5).

Table 5 Summary metrics for registered models seeking to explain deviation (Box-Cox transformed absolute deviation scores) from $\overline{Z}_r$ as a function of Sorensen's Index, categorical peer ratings, and continuous peer ratings for blue tit and *Eucalyptus* analyses, and as a function of the presence or absence of random effects (in the analyst's models) for *Eucalyptus* analyses. We report coefficient of determination, R^2, for our models including only fixed effects as predictors of deviation, and we report $R^2_{Conditional}$, $R^2_{Marginal}$ and the intra-class correlation (ICC) from our models that included both fixed and random effects. For all our models, we calculated the residual standard deviation σ and root mean squared error (RMSE)

Dataset	NObs	R^2	$R^2_{Conditional}$	$R^2_{Marginal}$	ICC	σ	RMSE
Deviation explained by categorical ratings							
Eucalyptus	346		0.13	0.01	0.12	1.06	1.02
blue tit	473		0.09	7.47×10^{-3}	0.08	0.5	0.48
Deviation explained by continuous ratings							
Eucalyptus	346		0.12	7.44×10^{-3}	0.11	1.06	1.03
blue tit	473		0.09	3.44×10^{-3}	0.09	0.5	0.48
Deviation explained by Sorensen's index							
Eucalyptus	79	1.84×10^{-4}				1.12	1.1
blue tit	131	6.32×10^{-3}				0.51	0.51
Deviation explained by inclusion of random effects							
Eucalyptus	79	8.75×10^{-8}				1.12	1.1

Table 6 Parameter estimates from models of Box-Cox transformed deviation scores from $\overline{Z}_r$ as a function of continuous and categorical peer ratings, Sorensen scores, and the inclusion of random effects. Standard errors (SE) and 95% confidence intervals (95% CI) are reported for all estimates, while *t* values, degrees of freedom and *p*-values are presented for fixed-effects. Note that positive parameter estimates mean that as the predictor variable increases, so does the absolute value of the deviation from the meta-analytic mean

Parameter	Random effect	Coefficient	SE	95% CI	t	*df*	*p*
Deviation explained by inclusion of random effects - *Eucalyptus*							
(Intercept)		−2.53	0.27	[−3.06, −1.99]	−9.31	77	<0.001
Mixed model		0.00	0.31	[−0.60, 0.60]	0.00	77	>0.9
Deviation explained by Sorensen's index - *Eucalyptus*							
(Intercept)		−2.65	1.05	[−4.70, −0.60]	−2.53	77	0.011
Mean Sorensen's index		0.18	1.51	[−2.78, 3.14]	0.12	77	>0.9
Deviation explained by Sorensen's index - blue tit							
(Intercept)		−1.53	0.28	[−2.08, −0.98]	−5.42	129	<0.001
Mean Sorensen's index		0.42	0.47	[−0.49, 1.34]	0.91	129	0.4
Deviation explained by continuous ratings - *Eucalyptus*							
(Intercept)		−2.23	0.23	[−2.69, −1.78]	−9.65	342	<0.001
RateAnalysis		−0.004	0	[−0.011, 0]	−1.44	342	0.15
SD(Intercept)	Reviewer ID	0.37	0.09	[0.24, 0.60]			
SD(Observations)	Residual	1.06	0.04	[0.98, 1.15]			
Deviation explained by continuous ratings - blue tit							
(Intercept)		−1.16	0.11	[−1.37, −0.94]	−10.60	469	<0.001
RateAnalysis		−0.002	0	[−0.004, 0]	−1.22	469	0.2
SD(Intercept)	Reviewer ID	0.16	0.03	[0.10,0.24]			
SD(Observations)	Residual	0.5	0.02	[0.46,0.53]			
Deviation explained by categorical ratings - *Eucalyptus*							
(Intercept)		−2.66	0.27	[−3.18, −2.13]	−9.97	340	<0.001
Publishable with major revision		0.29	0.29	[−0.27, 0.85]	1.02	340	0.3
Publishable with minor revision		0.01	0.28	[−0.54, 0.56]	0.04	340	>0.9
Publishable as is		0.05	0.31	[−0.55, 0.66]	0.17	340	0.9
SD(Intercept)	Reviewer ID	0.39	0.09	[0.25, 0.61]			
SD(Observations)	Residual	1.06	0.04	[0.98, 1.15]			
Deviation explained by categorical ratings - blue tit							
(Intercept)		−1.11	0.11	[−1.33, −0.89]	−9.91	467	<0.001
Publishable with major revision		−0.19	0.12	[−0.42, 0.04]	−1.62	467	0.10
Publishable with minor revision		−0.19	0.12	[−0.42, 0.04]	−1.65	467	0.10
Publishable as is		−0.13	0.13	[−0.39, 0.12]	−1.02	467	0.3
SD(Intercept)	Reviewer ID	0.15	0.04	[0.10, 0.24]			
SD(Observations)	Residual	0.5	0.02	[0.46, 0.53]			

Multivariate analysis effect size (Zr) and out-of-sample predictions (yi)

Like the univariate models, the multivariate models did a poor job of explaining deviations from the meta-analytic mean. Because we pre-registered a multivariate model that contained collinear predictors that produce results which are not readily interpretable, we present these models in the supplement. We also had difficulty with convergence and singularity for multivariate models of out-of-sample (y_i) result and had to adjust which random effects we included (Supplementary Material C, Table C.8). However, no multivariate analyses of *Eucalyptus* out-of-sample results avoided problems of convergence or singularity, no matter which random effects we included (Supplementary Material C, Table C.8). We therefore present no multivariate *Eucalyptus* y_i models. We present parameter estimates from multivariate Z_r models for both datasets (Supplementary

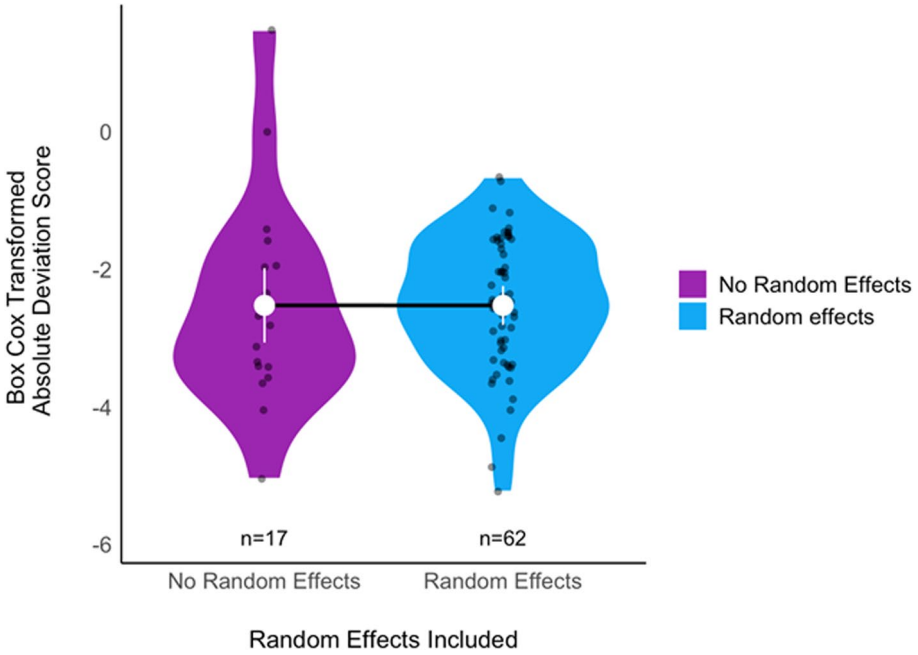


Fig. 6 Violin plot of mean Box-Cox transformed deviation from meta-analytic mean as a function of random-effects inclusion in *Eucalyptus* analyses. White point for each group of analyses denotes model-estimated marginal mean deviation, and error bars denote 95% CI of the estimate

Material C, Table6, Table7) and from y_i models from the blue tit dataset (Supplementary Material C, Table C.10, Table C.9). We include interpretation of the results from these models in the supplement, but the results do not change the interpretations we present above based on the univariate analyses.

Discussion

When a large pool of ecologists and evolutionary biologists analyzed the same two datasets to answer the corresponding two research questions, they produced substantially heterogeneous sets of answers. Although the variability in analytical outcomes was high for both datasets, the patterns of this variability differed distinctly between them. For the blue tit dataset, there was nearly continuous variability across a wide range of Z_r values. In contrast, for the *Eucalyptus* dataset, there was less variability across most of the range, but more striking outliers at the tails. Among out-of-sample predictions, there was again almost continuous variation across a wide range (2 SD) among blue tit estimates. For *Eucalyptus*, out-of-sample predictions were also notably variable, with about half the predicted stem count values at <2 but the other half being much larger, and ranging to nearly 40 stems per 15 m × 15 m plot. We investigated several hypotheses for drivers of this variability within datasets, but found little support for any of these. Most notably, even when we excluded analyses that had received one or more poor peer reviews, the heterogeneity in results largely persisted. Regardless of what drives the variability, the existence of such dramatically heterogeneous results when ecologists and evolutionary biologists seek to answer the same questions with the same data should trigger conversations about how ecologists and evolutionary biologists analyze data and interpret the results of their own analyses and those of others in the literature [8, 16, 96, 100].

Our observation of substantial heterogeneity due to analytical decisions is consistent with a small earlier study in ecology [103] and a growing body of work from the quantitative social sciences [15, 16, 23, 44, 92, 96]. In these studies, when volunteers from the discipline analyzed the same data, they produced a worryingly diverse set of answers to a pre-set question. This diversity included a wide range of effect sizes, and in most cases, even involved effects in opposite directions. Thus, our result should not be viewed as an anomalous outcome from two particular datasets, but instead as evidence

from additional disciplines regarding the heterogeneity that can emerge from analyses of complex datasets to answer questions in probabilistic science. Not only is our major observation consistent with other studies, it is, itself, robust because it derived primarily from simple forest plots that we produced based on a small set of decisions that were mostly registered before data gathering and which conform to widely accepted meta-analytic practices.

Unlike the strong pattern we observed in the forest plots, our other analyses, both registered and post hoc, produced either inconsistent patterns, weak patterns, or the absence of patterns. Our registered analyses found that deviations from the meta-analytic mean by individual effect sizes ($\overline{Zr}$) or the predicted values of the dependent variable ($\overline{y}$) were poorly explained by our hypothesized predictors: peer rating of each analysis team's method section, a measurement of the distinctiveness of the set of predictor variables included in each analysis, or whether the model included random effects. However, in our post hoc analyses, we found that dropping analyses identified as unpublishable or in need of major revision by at least one reviewer modestly reduced the observed heterogeneity among the Z_r outcomes, but only for *Eucalyptus* analyses, apparently because this led to the dropping of the major outlier. This limited role for peer review in explaining the variability in our results should be interpreted cautiously because the inter-rater reliability among peer reviewers was extremely low, and at least some analyses that appeared flawed to us were not marked as flawed by reviewers. Thus, it seems that the peer reviews we received were of mixed quality, possibly due to lack of expertise or lack of care on the part of some reviewers. However, the hypothesis that poor quality analyses drove a substantial portion of the heterogeneity we observed was also contradicted by our observation that analysts' self-declared statistical expertise appeared unrelated to heterogeneity. When we retained only analyses from teams including at least one member with high self-declared levels of expertise, heterogeneity among effect sizes remained high. Thus, our results suggest lack of statistical expertise is not the primary factor responsible for the heterogeneity we observed, although further work is merited before rejecting a role for statistical expertise. Besides variability in expertise, it is also possible that the volunteer analysts varied in the effort they invested, and low effort presumably drove at least some heterogeneity in results. However, analysts often submitted thoughtful and extensive code, tables, figures, and textual explanation and interpretations, which is evidence of substantial investment. Further, we are confident that low effort alone is an insufficient explanation for the heterogeneity we observed because we have worked with these datasets ourselves, and we know from experience that there are countless plausible modelling alternatives that can produce a diversity of effects. Additionally, heterogeneity in analytical outcomes differed notably between datasets, and there is no reason to expect that one set of analysts took this project less seriously than the other. Returning to our exploratory analyses, not surprisingly, simply dropping outlier values of Z_r for *Eucalyptus* analyses, which had more extreme outliers, led to less observable heterogeneity in the forest plots, and also reductions in our quantitative measures of heterogeneity. We did not observe a similar effect in the blue tit dataset because that dataset had outliers that were much less extreme and instead had more variability across the core of the distribution.

Our major observations raise two broad questions; why was the variability among results so high, and why did the pattern of variability differ between our two datasets. One important and plausible answer to the first question is that much of the heterogeneity derives from the lack of a precise relationship between the two biological research questions we posed and the data we provided. This lack of a precise relationship between data and question creates many opportunities for different model specifications, and so may inevitably lead to varied analytical outcomes [8]. However, we believe that the research questions we posed are consistent with the kinds of research question that ecologists and evolutionary biologists typically work from. When designing the two biological research questions, we deliberately sought to represent the level of specificity we typically see in these disciplines. This level of specificity is evident when we look at the research questions posed by some recent meta-analyses in these fields:

- "how [does] urbanization impact mean phenotypic values and phenotypic variation … [in] paired urban and non-urban comparisons of avian life-history traits" [22]
- "[what are] the effects of ocean acidification on the crustacean exoskeleton, assessing both exoskeletal ion content (calcium and magnesium) and functional properties (biomechanical resistance and cuticle thickness)" [95]
- "[what is] the extent to which restoration affects both the mean and variability of biodiversity outcomes … [in] terrestrial restoration" [7]
- "[does] drought stress [have] a negative, positive, or null effect on aphid fitness" [58]

- "[what is] the influence of nitrogen-fixing trees on soil nitrous oxide emissions" [53]

There is not a single precise answer to any of these questions, nor to the questions we posed to analysts in our study. And this lack of single clear answers will obviously continue to cause uncertainty since ecologists and evolutionary biologists conceive of the different answers from the different statistical models as all being answers to the same general question. A possible response would be a call to avoid these general questions in favor of much more precise alternatives [8]. However, the research community rewards researchers who pose broad questions [98], and so researchers are unlikely to narrow their scope without a change in incentives. Further, we suspect that even if individual studies specified narrow research questions, other scientists would group these more narrow questions into broader categories, for instance in meta-analyses, because it is these broader and more general questions that often interest the research community.

Although variability in statistical outcomes among analysts may be inevitable, our results raise questions about why this variability differed between our two datasets. We are particularly interested in the differences in the distribution of Z_r since the distributions of out-of-sample predictions were on different scales for the two datasets, thus limiting the value of comparisons. The forest plots of Z_r from our two datasets showed distinct patterns, and these differences are consistent with several alternative hypotheses. The results submitted by analysts of the *Eucalyptus* dataset showed a small average (close to zero) with most estimates also close to zero (± 0.2), though about a third far enough above or below zero to cross the traditional threshold of statistical significance. There were a small number of striking outliers that were very far from zero. In contrast, the results submitted by analysts of the blue tit dataset showed an average much further from zero (−0.35) and a much greater spread in the core distribution of estimates across the range of Z_r values (± 0.5 from the mean), with few modest outliers. So, why was there more spread in effect sizes (across the estimates that are not outliers) in the blue tit analyses relative to the *Eucalyptus* analyses?

One possible explanation for the lower heterogeneity among most *Eucalyptus* Z_r effects is that weak relationships may limit the opportunities for heterogeneity in analytical outcome. Some evidence for this idea comes from two sets of "many labs" studies in psychology [49, 50]. In these studies, many independent lab groups each replicated a large set of studies, including, for each study, the experiment, data collection, and statistical analyses. These studies showed that, when the meta-analytic mean across the replications from different labs was small, there was much less heterogeneity among the outcomes than when the mean effect sizes were large [49, 50]. Of course, a weak average effect size would not prevent divergent effects in all circumstances. As we saw with the *Eucalyptus* analyses, taking a radically smaller subset of the data can lead to dramatically divergent effect sizes even when the mean with the full dataset is close to zero.

Our observation that dramatic sub-setting in the *Eucalyptus* dataset was associated with correspondingly dramatic divergence in effect sizes leads us towards another hypothesis to explain the differences in heterogeneity between the *Eucalyptus* and blue tit analysis sets. It may be that when analysts often divide a dataset into subsets, the result will be greater heterogeneity in analytical outcome for that dataset. Although we saw sub-setting associated with dramatic outliers in the *Eucalyptus* dataset, nearly all other analyses of *Eucalyptus* data used close to the same set of 351 samples, and as we saw, these effects did not vary substantially. However, analysts often analyzed only a subset of the blue tit data, and as we observed, sample sizes were much more variable among blue tit effects, and the effects themselves were also much more variable. Important to note here is that subsets of data may differ from each other for biological reasons, but they may also differ due to sampling error. Sampling error is a function of sample size, and sub-samples are, by definition, smaller samples, and so more subject to variability in effects due to sampling error [46].

Other features of datasets are also plausible candidates for driving heterogeneity in analytical outcomes, including features of covariates. In particular, relationships between covariates and the response variable as well as relationships between covariates and the primary independent variable (collinearity) can strongly influence the modeled relationship between the independent variable of interest and the dependent variable [27, 70]. Therefore, inclusion or exclusion of these covariates can drive heterogeneity in effect sizes (Z_r). Also, as we saw with the two most extreme Z_r values from the blue tit analyses, in multivariate models with collinear predictors, extreme effects can emerge when estimating partial correlation coefficients due to high collinearity, and conclusions can differ dramatically depending on which relationship receives the researcher's attention. Therefore, differences between datasets in the presence of strong and/

or collinear covariates could influence the differences in heterogeneity in results among those datasets.

Although it is too early in the many-analyst research program to conclude which analytical decisions or which features of datasets are the most important drivers of heterogeneity in analytical outcomes, we must still grapple with the possibility that analytical outcomes may vary substantially based on the choices we make as analysts. If we assume that, at least sometimes, different analysts will produce dramatically different statistical outcomes, what should we do as ecologists and evolutionary biologists? We review some ideas below.

The easiest path forward after learning about this analytical heterogeneity would be simply to continue with "business as usual", where researchers report results from a small number of statistical models. A case could be made for this path based on our results. For instance, among the blue tit analyses, the precise values of the estimated Z_r effects varied substantially, but the average effect was convincingly different from zero, and a majority of individual effects (84%) were in the same direction. Arguably, many ecologists and evolutionary biologists appear primarily interested in the direction of a given effect and the corresponding *p*-value [30], and so the variability we observed when analyzing the blue tit dataset may not worry these researchers. Similarly, most effects from the *Eucalyptus* analyses were relatively close to zero, and about two-thirds of these effects did not cross the traditional threshold of statistical significance. Therefore, a large proportion of people analyzing these data would conclude that there was no effect, and this is consistent with what we might conclude from the meta-analysis.

However, we find the counter arguments to "business as usual" to be compelling. For blue tits, there were a substantial minority of calculated effects that would be interpreted by many biologists as indicating the absence of an effect (28%), and there were three traditionally "significant" effects in the opposite direction to the average. The qualitative conclusions of analysts also reflected substantial variability, with fully half of teams drawing a conclusion distinct from the one we draw from the distribution as a whole. These teams with different conclusions were either uncertain about the negative relationship between competition and nestling growth, or they concluded that effects were mixed or absent. For the *Eucalyptus* analyses, this issue is more concerning. Around two-thirds of effects had confidence intervals overlapping zero, and of the third of analyses with confidence intervals excluding zero, almost half were positive, and the rest were negative. Accordingly, the qualitative conclusions of the *Eucalyptus* teams were spread across the full range of possibilities. But, as we describe in the next paragraph, even this striking lack of consensus may be much less of a problem than what could emerge as scientists select which results to publish.

A potentially larger argument against "business as usual" is that it provides the raw material for biasing the literature. When different model specifications readily lead to different results, analysts may be tempted to report the result that appears most interesting, or that is most consistent with expectation [36, 33]. There is growing evidence that researchers in ecology and evolutionary biology often report a biased subset of the results they produce [26, 48] and that this bias exaggerates the average size of effects in the published literature between 30 and 150% [83, 121]. The bias then accumulates in meta-analyses, apparently more than doubling the rate of conclusions of "statistical significance" in published meta-analyses above what would have been found in the absence of bias [121]. Thus, "business as usual" does not just create noisy results, it helps create systematically misleading results.

If we move away from "business as usual", where do we go? Many obvious options involve multiple analyses per dataset. For instance, there is the traditional robustness or sensitivity check [17, 85], in which the researcher presents several alternative versions of an analysis to demonstrate that the result is "robust" [59]. Unfortunately, robustness checks are at risk of the same potential biases of reporting found in other studies [96], especially given the relatively few models typically presented. However, these risks could be minimized by running more models and doing so with a pre-registration or registered report. Another option is model averaging. Averages across models often perform well [105], and in some forms this may be a relatively simple solution. Model averaging, as most often practiced in ecology and evolutionary biology, involves first identifying a small suite of candidate models [20], then using Akaike weights, baBaselgased on Akaike's Information Criterion (AIC), to calculate weighted averages for parameter estimates from those models. As with typical robustness checks, the small number of models limits the exploration of specification space, but examining a larger number of models could become the norm. However, there are more concerning limitations. The largest of these limitations is that averaging regression coefficients is problematic when models differ in interaction terms or collinear variables [21]. Additionally, weighting by AIC may often be inconsistent with our

modelling goals. AIC balances the trade-off between model complexity and predictive ability, but penalizing models for complexity may not be suited for testing hypotheses about causation [4]. So, AIC may often not offer the weight we want to use, and we may also not wish to just generate an average at all. Instead, if we hope to understand an extensive universe of possible modelling outcomes, we could conduct a multiverse analysis, possibly with a specification curve [99, 100]. This could mean running hundreds or thousands of models (or more!) to examine the distribution of possible effects, and to see how different model specification choices map onto these effects. However, exploring large areas of specification space may come at the cost of including biologically implausible specifications. Thus, we expect a trade-off, and attempts to limit models to the most biologically plausible may become increasingly difficult in proportion to the number of variables and modelling choices. To make selecting plausible models easier, one could recruit multiple analysts to design one or a few plausible specifications each as with our "many analyst" study [96]. An alternative that may be more labor intensive for the primary analyst, but which may lead to a more plausible set of models, could involve hypothesizing about causal pathways with DAGs [directed acyclic graphs, Arif and MacNeil ([5])] to constrain the model set. As with other options outlined above, generating model specifications with DAGs could be partnered with pre-registration to hinder bias from undisclosed data dredging.

Responses to heterogeneity in analysis outcomes need not be limited to simply conducting more analyses, especially if it turns out that analysis quality drives some of the observed heterogeneity. As we noted above, we cannot yet rule out the possibility that insufficient statistical expertise or poor-quality analyses might drive some portion of the heterogeneity we observed. Improving the quality of analyses might be accomplished with a deliberate increase in investment in statistical education. Many ecology and evolutionary biology students learn their statistical practice informally, with many ecology doctoral programs in the USA not requiring a statistics course [107]), and no formal courses of any kind included in doctoral degrees in most other countries. In cases where formal investment in statistical education is lacking, informal resources, such as guidelines and checklists, may help researchers avoid common mistakes. However, unless following guidelines or checklists is enforced for publication, the adherence to guidelines is patchy. For example, despite the publication of guidelines for conducting meta-analyses in ecology, the quality of meta-analyses did not improve substantially over time [52]. Even in medical research where adherence to guidelines such as the PRISMA standards for systematic reviews and meta-analyses is more highly valued, adherence is often poor [81].

Although we have reviewed a variety of potential responses to the existence of variability in analytical outcomes, we certainly do not wish to imply that this is a comprehensive set of possible responses. Nor do we wish to imply that the opinions we have expressed about these options are correct. Determining how the disciplines of ecology and evolutionary biology should respond to knowledge of the variability in analytical outcome will benefit from the contribution and discussion of ideas from across these disciplines. We look forward to learning from these discussions and to seeing how these disciplines ultimately respond.

Conclusions

Overall, our results suggest to us that, where there is a diverse set of plausible analysis options, no single analysis should be considered a complete or reliable answer to a research question. Further, because of the evidence that ecologists and evolutionary biologists often present a biased subset of the analyses they conduct [26, 48, 121], we do not expect that even a collection of different effect sizes from different studies will accurately represent the true distribution of effects [121]. Therefore, we believe that an increased level of skepticism of the outcomes of single analyses, or even single meta-analyses, is warranted going forward. We recognize that some researchers have long maintained a healthy level of skepticism of individual studies as part of sound and practical scientific practice, and it is possible that those researchers will be neither surprised nor concerned by our results. However, we doubt that many researchers are sufficiently aware of the potential problems of analytical flexibility to be appropriately skeptical. We hope that our work leads to conversations in ecology, evolutionary biology, and other disciplines about how best to contend with heterogeneity in results that is attributable to analytical decisions.

Appendix 1

R Package References and Session Information

Table 7 shows all R packages and their versions used in the production of the manuscript.

Table 7 R packages used to generate this manuscript. Please see the ManyEcoEvo: package for a full list of packages used in the analysis pipeline

Package	Version	Citation
base	4.4.0	R Core Team (R Core Team [87]
betapart	1.6	Baselga et al. (2023)
broom.mixed	0.2.9.5	Bolker et al. [13]
colorspace	2.1.0	Zeileis et al. [122]
cowplot	1.1.3	Wilke [118]
devtools	2.4.5	Wickham et al. (Wickham et al. [116])
EnvStats	2.8.1	Millard (Millard [68])
GGally	2.2.1	Schloerke et al. (Schloerke et al. [91])
ggforestplot	0.1.0	Scheinin et al. (Scheinin et al. [90])
ggh4x	0.2.8	van den Brand [109]
ggpubr	0.6.0	Kassambara (Kassambara [47])
ggrepel	0.9.5	Slowikowski (Slowikowski [102])
ggthemes	5.1.0	Arnold [6]
glmmTMB	1.1.8	Brooks et al. ([18])
gt	0.10.1	Iannone et al. (Iannone et al. [45])
gtsummary	1.7.2	Sjoberg et al. [101]
here	1.0.1	Müller Müller [71]
Hmisc	5.1.2	Harrell Jr [41]
irr	0.84.1	Gamer, Lemon, and Singh ([35])
janitor	2.2.0	Firke (Firke [32])
knitr	1.46	Xie (Xie [119])
latex2exp	0.9.6	Meschiari (Meschiari [66])
lme4	1.1.35.3	Bates et al. [11]
ManyEcoEvo	2.7.6	Gould et al. [38]
metafor	4.6.0	Viechtbauer [114]
modelbased	0.8.7	Makowski et al. (Makowski et al. [64])
multilevelmod	1.0.0	Kuhn and Frick (Kuhn and Frick [54])
MuMIn	1.47.5	Bartoń [9]
naniar	1.1.0	Tierney and Cook [106]
NatParksPalettes	0.2.0	Blake (Blake [12])
orchaRd	2	Nakagawa, Lagisz, et al. (2023)
parameters	0.21.7	Lüdecke et al. [60]
patchwork	1.2.0	Pedersen [84]
performance	0.11.0	Lüdecke, Ben-Shachar, et al. [61]
renv	1.0.2	Ushey and Wickham (Ushey and Wickham [108])
rmarkdown	2.27	Allaire et al. [3]
sae	1.3	Molina and Marhuenda (Molina and Marhuenda [69])
scales	1.3.0	Wickham, Pedersen, and Seidel [117]
see	0.8.4	Lüdecke, Patil, et al. [62]
showtext	0.9.7	Qiu Qiu [86]
specr	1.0.0	Masur and Scharkow (Masur and Scharkow [65])
targets	1.7.0	Landau [57]
tidymodels	1.1.1	Kuhn and Wickham [55]
tidytext	0.4.2	Silge and Robinson [97]
tidyverse	2.0.0	Wickham et al. [115]
withr	3.0.0	Hester et al. (Hester et al. [42])
xfun	0.44	Xie (Xie [120])

Appendix 2

Table 8 Session info

setting	value
version	R version 4.4.0 (2024-04-24)
os	macOS Ventura 13.6.9
system	aarch64, darwin20
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Australia/Melbourne
date	2024-09-17
pandoc	3.1.12.2 @ /opt/homebrew/bin/ (via rmarkdown)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-02101-x.

Supplementary Material 1.

Authors' contributions

HF, THP and FF conceptualized the project. PV provided raw data for *Eucalyptus* analyses and SG and THP provided raw data for blue tit analyses. DGH, HF and THP prepared surveys for collecting participating analysts and reviewer's data. EG, HF, THP, PV, SN and FF planned the analyses of the data provided by our analysts and reviewers, EG, HF, DGH and THP curated the data, EG and HF wrote the software code to implement the analyses and prepare data visualizations. EG ensured that analyses were documented and reproducible. THP and HF administered the project, including coordinating with analysts and reviewers. FF provided funding for the project. THP, HF, and EG wrote the manuscript. Authors listed alphabetically contributed analyses of the primary datasets or reviews of analyses. All authors read and approved the final manuscript.

Funding

EG's contributions were supported by an Australian Government Research Training Program Scholarship, AIMOS top-up scholarship (2022) and Melbourne Centre of Data Science Doctoral Academy Fellowship (2021). FF's contributions were supported by ARC Future Fellowship FT150100297.

Data availability

All materials and data are archived and hosted on the OSF at https://osf.io/mn5aj/, including survey instruments and analyst / reviewer consent forms. The Evolutionary Ecology Data and Ecology and Conservation Data provided to analysts are available at https://osf.io/34fzc/ and https://osf.io/t76uy/ respectively. Data has been anonymized, and the non-anonymized data is stored on the project OSF within private components accessible to the lead authors.

We built an R package, ManyEcoEvo to conduct the analyses described in this study [38], which can be downloaded from https://github.com/egouldo/ManyEcoEvo/ to reproduce our analyses or replicate the analyses described here using alternate datasets. Data cleaning and preparation of analysis-data, as well as the analysis, is conducted in [87] reproducibly using the targets package [57]. This data and analysis pipeline is stored in the ManyEcoEvo package repository and its outputs are made available to users of the package when the library is loaded.

The full manuscript, including further analysis and presentation of results is written in Quarto [2]. The source code to reproduce the manuscript is hosted at https://github.com/egouldo/ManyAnalysts/, and the rendered version of the source code may be viewed at https://egouldo.github.io/ManyAnalysts/. All R packages and their versions used in the production of this manuscript are listed in Appendix 1.

Declarations

Ethics approval and consent to particpate

We obtained permission to conduct this research from the Whitman College Institutional Review Board (IRB). As part of this permission, the IRB approved the consent form (https://osf.io/xyp68/) that all participants completed prior to joining the study. The authors declare that they have no competing interests.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Agriculture Food and Ecosystem Sciences, University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia. [2]School of Historical and Philosophical Studies, University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia. [3]Department of Biology, Whitman College, 345 Boyer Ave, Walla Walla, WA 99362, USA. [4]School of Biological, Earth & Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia. [5]School of Natural Sciences, Macquarie University, Balaclava Rd, Macquarie Park, Sydney, NSW 2109, Australia. [6]School of Public Health and Preventive Medicine, Monash University, 750 Collins Street, Docklands, VIC 3008, Australia. [7]Länsstyrelsen Östergötland, Östgötagatan 3, 58186 Linköping, Sweden. [8]Biology Department, Lund University, Sölvegatan 37, 22362 Lund, Sweden. [9]Department of Biology, University of Massachusetts, 1 Campus Center Way, Amherst, MA 01003, USA. [10]Marine and Continental Waters, IRTA, Carretera Poble Nou Km 5.5, 43540 La Ràpita, Catalonia, Spain. [11]Department of Life Sciences, Ben Gurion University of the Negev, P.O.Box 653, 84105 Beer Sheva, Israel. [12]Department of Psychology, The University of Edinburgh, 7 George Square, Edinburgh EH9 1HB, UK. [13]Centre for Ecological Sciences, Indian Institute of Science, Indian Institute of Science, Bengaluru, Karnataka 560012, India. [14]Southern Research Station, USDA Forest Service, PO Box 700, New Ellenton, SC 29809, USA. [15]Center for Ecological Dynamics in a Novel Biosphere (ECONOVO), Department of Biology, Aarhus University, Ny Munkegade 114-116, 8000 Aarhus C, Denmark. [16]School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3052, Australia. [17]Biology, Indiana University Purdue University Indianapolis, 420 University Blvd, Indianapolis, IN 46202, USA. [18]School of Life and Environmental Sciences, Deakin University, 221 Burwood Highway, Burwood, VIC 3125, Australia. [19]Department of Arid Land Agriculture, King Abdulaziz University, Jeddah 80200, Kingdom of Saudi Arabia. [20]Department of Biological Sciences, Macquarie University, 205ACR Culloden Road, Macquarie Park, New South Wales 2113, Australia. [21]Department of Plant Ecology, University of Hohenheim, Institute of Landscape and Plant Ecology, Ottilie-Zeller-Weg, 70599 Stuttgart, Germany. [22]Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Skogsmarksgränd 17, SE-907 36 Umeå, Sweden. [23]Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Rd, Storrs, CT 06226, USA. [24]STEM Center, Southern Illinois University Edwardsville, 1 Hairpin Dr, Edwardsville, IL 62026, USA. [25]University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada. [26]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland. [27]Department of Biological Sciences, California State Polytechnic University, Pomona, USA. [28]Centre d'Études Biologiques de Chizé, UMR 7372, Université de la Rochelle - Centre National de la Recherche Scientifique, 405 route de Prissé la Charrière, 79360 Villiers en Bois, France. [29]Faculty of Life Sciences, Bar Ilan University, Ramat Gan 529000, Israel. [30]School of Natural Sciences, University of Tasmania, TAS, Private Bag 55, Hobart 7001, Australia. [31]Whitebark Institute, 3399 Main Street, Suite W5, Mammoth Lakes, CA 93546, USA. [32]Department of Biology, Rhodes College, 2000 N., Parkway, Memphis, TN 38112, USA. [33]Centre for Conservation Science, RSPB, 2 Lochside View, Edinburgh EH12 9DH, UK. [34]Environmental Studies, Wofford College, 429 N. Church St, Spartanburg, SC 29303, USA. [35]IFM Biology, Linköping University, 581 83 Linköping, Sweden. [36]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Czech Republic, Kamýcká 129, Praha

- Suchdol 165 00, Czech Republic. [37]Biological and Environmental Sciences & Gothenburg Global Biodiversity Centre, University of Gothenburg, Medicinaregatan 7B, SE-413 90 Gothenburg, Sweden. [38]School of Biological Sciences, Monash University, Rainforest Walk 25, Clayton, Victoria, Australia. [39]Ecology and Evolutionary Biology, University of Tennessee Knoxville, 569 Dabney Hall, Knoxville, TN 37996, USA. [40]Departamento de Ecologia e Zoologia, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, UFSC, Campus Universitário - Córrego Grande Florianópolis – SC; CEP, Florianópolis 88040-900, Brazil. [41]School of Biological and Forensic Sciences, University of South Wales, The Alfred Russel Wallace Building, 9 Graig Fach, Glyntaff, Pontypridd CF37 4BB, UK. [42]Centre for Ecology, Evolution and Environmental Changes (cE3c) &, CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal. [43]Forest and Rangeland Stewardship, Colorado State University, 1472 Campus Delivery, Fort Collins, CO 80523-1472, USA. [44]Department of Ornithology, Academy of Natural Sciences of Drexel University, 1900 Benjamin Franklin Parkway, Philadelphia, PA 19096, USA. [45]Salisbury University, 1101 Camden Ave, Biology, Salisbury, MD 21801, USA. [46]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Nijenborgh 7, 9747 AG Groningen, Netherlands. [47]Department of Biology, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada. [48]Department of Zoology, University of Cambridge, Downing St, Cambridge CB2 3EJ, UK. [49]Entomology and Nematology, University of Florida, 700 Experiment Station Rd, Lake Alfred, FL 33850, USA. [50]Environmental Studies, Elon University, McMichael Science Building, 2625 Campus Box, Elon, NC 27244, USA. [51]BirdLife International, David Attenborough Building, Pembroke Street, Cambridge CB2 3QZ, UK. [52]School of Integrative Biological and Chemical Sciences, The University of Texas Rio Grande Valley, One West University Boulevard, Brownsville, TX 78520, USA. [53]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks St, Toronto, ON M5S 3B2, Canada. [54]Department of Statistics, University of Auckland, Auckland, New Zealand. [55]LLC, Catbird Stats, PO Box 2018, Gautier, MS 39553, USA. [56]School of Biodiversity, One Health & Veterinary Medicine, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK. [57]School of Forestry, Northern Arizona University, 200 E Pine Knoll Dr., Flagstaff, AZ 86001, USA. [58]Department of Behavioural Neurobiology, Max Planck Institute for Biological Intelligence, Eberhard-Gwinner-Strasse, 82319 Seewiesen, Oberbayern, Germany. [59]School of Sciences: Center for Health and Cognition, Bath Spa University, Newton Park, Bath BA2 9BN, UK. [60]Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2R3, Canada. [61]Applied Zoology, Zellescher Weg 20b, 01217 Dresden, TUDresden,, Germany. [62]School of Molecular Biosciences, College of Medical Veterinary & Life Sciences, University of Glasgow, Glasgow G12 8Qq, UK. [63]School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW 2006, Australia. [64]Institute of Environmental Sciences, Jagiellonian University, Gronostajowa 7, 30-387 Krakow, Poland. [65]Biology Department, Brigham Young University, 4102 Life Science Building, Provo, UT, USA. [66]Institute of Evolutionary Sciences Montpellier, University of Montpellier, CNRS, IRD, Montpellier, France. [67]Baruch Marine Field Laboratory, University of South Carolina, 2306 Crabhaul Rd, Georgetown, SC 29440, USA. [68]Department of Life Sciences, Imperial College London, Buckhurst road, Berkshire SL5 7PY, UK. [69]European Forest Institute, Platz d. Vereinten Nationen 7, 53113 Bonn, Germany. [70]Department of Ornithology, Max Planck Institute for Biological Intelligence, Eberhard-Gwinner Str. 7, 82319 Seewiesen, Germany. [71]Forestry and Environmental Conservation, National Bobwhite and Grassland Initiative, Clemson University, 243 Lehotsky Hall, Clemson, SC 29634, USA. [72]Department of Psychology and Vision Science, University of Birmingham, 52 Pritchatts Road. Edgbaston, Baily Thomas GrantBirmingham B15 2TT, UK. [73]School of Biology and Environmental Science, University College Dublin, Dublin 4, Belfield D04 V1W8, Ireland. [74]Department of Biological Sciences, San José State University, 129 S 10th Street, San Jose, CA 95112, USA. [75]Apicultural State Institute, University of Hohenheim, Erna-Hruschka-Weg 6, 70599 Stuttgart, Germany. [76]Department of Biology, St. Norbert College, 100 Grant St, De Pere, WI 54115, USA. [77]Department of Biological Sciences, Purdue University, 915 W. State Street, West Lafayette, IN 47907, USA. [78]Biodiversity, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Macroecology & BiogeographyBüsgenweg 1, 37077 Göttingen, Germany. [79]Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota-Twin Cities, 135 Skok Hall, 2003 Upper Buford Circle, St. Paul, MN 55108, USA. [80]CABI, Bakeham Lane, Egham, Surrey, UK. [81]Department of Ecology and Evolutionary Biology, School of Biological Sciences, University of California, 321 Steinhaus Hall, IrvineIrvine, CA 92697, USA. [82]School of Biological Sciences, University of Aberdeen, King Street, Aberdeen AB244FX, UK. [83]Department of Natural Resource Sciences, McGill University, 21111 Lakeshore Rd, Ste Anne-de-Bellevue, Montreal, QC H9X 3V9, Canada. [84]Institute of Ecology and Environmental Sciences (iEES), Univ. Paris-Est Creteil, 61 avenue du Général de Gaulle, 94010 Créteil, France. [85]Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, OR 97331, USA. [86]Wake Forest University, 1834 Wake Forest Road, Winston Salem, NC 27109, USA. [87]Laboratorio de Invasiones Biológicas (LIB), Instituto de Ecología y Biodiversidad, Victoria 631, Concepción, Chile. [88]Institute for Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow G12 8QQ, UK. [89]Department of Forest Ecosystems and Society, College of Forestry, Oregon State University, Corvallis, OR 97333, USA. [90]Biodiversity and Conservation Area, Rey Juan Carlos University, C/ Tulipán s/n, 28933 Móstoles, Madrid, Spain. [91]Graduate School of Biomedical Sciences, Tufts University, 136 Harrison Ave #813, Boston, MA 02111, USA. [92]Department of Integrative Biology, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada. [93]School of Life and Environmental Sciences (Burwood Campus), Deakin University, Geelong, Victoria, Australia. [94]CNRS, University of Rennes, 263 Avenue du Général Leclerc, 35042 Rennes, France. [95]Department of Behavioural Ecology, Bielefeld University, Konsequenz 45, 33615 Bielefeld, Germany. [96]Fakultät für Biologie, Arbeitsgruppe Evolutionsbiologie, Universität Bielefeld, Morgenbreede 45, 33615 Bielefeld, Germany. [97]Chair of Meteorology, Institute for Hydrology and Meteorology, Faculty of Environmental Sciences, Technische Universität Dresden, Pienner Str. 23, 01737 Tharandt, Germany. [98]Department of Wildlife, Fisheries, and Conservation Biology, University of Maine, 5755 Nutting Hall, Room 210, Orono, ME 04469-5755, USA. [99]Department of Biological Sciences, Boise State University, 1910 W University Dr, Boise, ID 83725, USA. [100]School of Agriculture, Food and Ecosystem Sciences, The University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia. [101]Pastures Systems and Watershed Management Research Unit, USDA Agricultural Research Service, USDA-ARS PSWMRU, Bldg. 3702 Curtin Road, University Park, PA 16802, USA. [102]Department of Natural Sciences, Baruch College, City University of New York, 17 Lexington Avenue, New York, NY 10010, USA. [103]Department of Biological Sciences, Binghamton University, 4400 Vestal Parkway East, Binghamton, NY 13902, USA. [104]Dipartimento di Biologia, Università di Roma "Tor Vergata", Via Cracovia, 1, 00133 Rome, Italy. [105]Department of Anthropology, University of Michigan, 1085 S. University Ave, Ann Arbor, MI 48109, USA. [106]College of Forestry, Oregon State University, 3100 SW Jefferson Way, Corvallis, OR 97333, USA. [107]University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, België, Belgium. [108]Earth & Environmental Sciences, Wesleyan University, 45 Wyllys Ave, Middletown, CT 06459, USA. [109]Department of Psychiatry, Yale School of Medicine, Yale University, 389 Whitney Ave, New Haven, CT 06511, USA. [110]Biology Department and Environmental Studies Department, St. Olaf College, 1520 St Olaf Ave, Northfield, MN 55057, USA. [111]Department of Plant Protection, Faculty of Agriculture, Department of Plant Protection, Faculty of Agriculture, Ordu University, Ordu University, 52200 Altinordu/Ordu, Turkey. [112]Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001 Leuven, Belgium. [113]Department of Marine Sciences, University of Gothenburg, Box 461, SE-40530 Gothenburg, Sweden. [114]Department of Biology, Wilfrid Laurier University, 75 University Ave West, Waterloo, Ontario N2L 3C5, Canada. [115]Natural Resource Ecology and Management, Iowa State University, 2310 Pammel Dr, Ames, IA 50011, USA. [116]School of Biological Sciences, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia. [117]Department of Biological Sciences, Middle East Technical University, Üniversiteler Mahallesi, Dumlupınar Bulvarı No: 1, 06800 Çankaya/Ankara, Turkey. [118]Dept. of Integrative Biology, University of Texas at Austin,2415 Speedway #C0930,, Austin, TX, USA. [119]Grand Bay National Estuarine Research Reserve, 6005 Bayou Heron Rd, Moss Point, MS 39562, USA. [120]Universidad de los Andes, Carrera 1 # 18A-12, Bogotá, Colombia. [121]Department of Biology, University of Turku, Turun Yliopisto, FI-20014 Turku, Finland. [122]Instituto de Ciencia Animal. Facultad de Ciencias Veterinarias, Universidad Austral de Chile, Campus Isla Teja s/n, Valdivia, Chile. [123]Department of Integrative Biology, The University of Texas at Austin, 2415 Speedway #C0930, Austin, Texas 78712, USA. [124]Department of Botany, University of Wyoming, Laramie, WY 82071, USA. [125]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden. [126]Department of Biology, Brigham Young University, Brigham Young University, Brigham Young University, Provo, UT 84602, USA. [127]School of Biological Sciences, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK. [128]International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1 , A-2361 Laxenburg, Austria. [129]Department of Ecology,

Swedish University of Agricultural Sciences, Ulls Väg 16, 750 07 Uppsala, Sweden. [130]Université de Montpellier, ISEM, University of Montpellier, CNRS, EPHE, 34000 Montpellier, IRD, France. [131]Department of Wildlife, Fish, and Conservation Biology, University of California, 1 Shields Ave, DavisDavis, CA 95616, USA. [132]Département des Sciences biologiques, Université du Québec à Montréal, 141 Avenue du Président-Kennedy, Montréal, Québec H2X 1Y4, Canada. [133]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JW, UK. [134]Center for Limnology, University of Wisconsin - Madison, 680 N Park St, Madison, WI 53706, USA. [135]Center for Biodiversity and Global Change, Yale University, 165 Prospect St, New Haven, CT 06511, USA. [136]Department of Ecology, Evolution, and Behavior, University of Minnesota, Ecology Building, 1987 Upper Buford Cir, St. PaulSt Paul, MN 55108, USA. [137]Institute of Environment and Department of Biological Sciences, Florida International University, 3000 NE 151st St, North Miami, FL 33181, USA. [138]Department of Life Sciences, Aberystwyth University, Penglais, Aberystwyth SY23 3DA, UK. [139]Institut de recherche en biologie végétale, Université de Montréal, 4101, Sherbrooke St E, Montréal, Québec H1X 2B2, Canada. [140]Institute of Evolutionary Ecology and Conservation Genomics, Ulm University, Albert-Einstein-Allee 11, 89081 Ulm, Germany. [141]Department of Biology, Memorial University of Newfoundland, 45 Arctic Ave, St John's NL A1C5S7, Canada. [142]Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, UNSW Sydney, High Street 2052, Kensington, NSW, Australia. [143]The Nature Conservancy, 258 Main Street, Lander, WY 82520, USA. [144]Ecology and Evolutionary Biology, University of Arizona, 1041 E Lowell St, Tucson, AZ 85721, USA. [145]Biological and Environmental Sciences, University of Stirling, Cottrell Building, Stirling FK9 4LA, UK. [146]Department of Biology, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium. [147]Aquatic Ecology and Evolution Group, Limnological Institute, University of Konstanz, Mainaustraße 252, 78464 Konstanz, Germany. [148]Campus Cerro Largo, Universidade Federal da Fronteira Sul, Rua Jacob Haupenthal, Cerro Largo, RS, CEP 158097900-000, Brazil. [149]School of Psychology and Social Work, University of Hull, Cottingham Rd, Hull HU6 7RX, UK. [150]UMR 1224, ECOBIOP, Université de Pau et des Pays de l'Adour, 173 Route de Saint-Jean-de-Luz, 64310 Saint-Pée-sur-Nivelle, France. [151]School of Biological & Environmental Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool L3 3AF, UK. [152]Institute of Ecology and Evolution, University of Edinburgh, The University of Edinburgh, King's Buildings, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK. [153]Phnom Penh, Cambodia. [154]Statistical Ecotoxicology, Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Universitätsstraße 30, 95440 Bayreuth, Germany. [155]Ecology and Environmental Science, Umeå University, Linnaeus väg 6, 907 36 Umeå, Sweden. [156]Molecular Ecology Group (MEG), Water Research Institute (IRSA), National Research Council of Italy (CNR), 28922 Corso Tonolli 50, Verbania, Italy. [157]Department of Natural History, Norwegian University of Science and Technology, Høgskoleringen 1, 7034 Trondheim, Norway. [158]Center for Advanced Biotechnology and Medicine, Rutgers University Robert Wood Johnson Medical School, 679 Hoes Lane West, Piscataway, NJ 08854, USA. [159]Departamento de Ciencias Biológicas, Universidad de los Andes, Carrera 1 Nº 18A - 12, 111711 Bogotá, Bogotá D. C, Colombia. [160]Institut de recherche sur les forêts, Université du Québec en Abitibi-Témiscamingue, 445 Boulevard de l'Université, Rouyn-Noranda, QC J9X 5E4, Canada. [161]Université du Québec à Trois-Rivières, 3351, boulevard des Forges, Trois-Rivières (Québec) G8Z 4M3, Canada. [162]Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland. [163]School of Biological, Earth and Environmental Sciences, University of New South Wales, Randwick, Sydney, NSW 2052, Australia. [164]Whitney Laboratory for Marine Bioscience, University of Florida, 9505 N Ocean Shore Blvd, St. Augustine, Gainesville, FL 32080, USA. [165]Biological Sciences, Eastern Illinois University, 600 Lincoln Avenue, Charleston, IL 61920, USA. [166]Centre d'Investigations Clinique Plurithématique - Institut Lorrain du Coeur et des Vaisseaux, Université de Lorraine, Inserm1433 CIC-P CHRU de Nancy, bâtiment Louis Mathieu - 5, rue du Morvan - 54500, Vandoeuvre-les-nancy, France. [167]Evolutionary biology department, Bielefeld University, Bielefeld University, Konsequenz 45, 33615 Bielefeld, Germany. [168]Department of Ecology and global change, Centro de Investigaciones sobre Desertificación, Consejo Superior de Investigaciones Cientificas (CIDE-CSIC/UV/GV), Carretera CV-315 km 10,7, 46113 Moncada (Valencia), Spain. [169]Department of Biological Sciences, University of Arkansas, 850 W. Dickson Street SCEN601, Fayetteville, AR 72701, USA. [170]School of the Environment, Faculty of Science, The University of Queensland, The University of Queensland, Brisbane, QLD 4072, Australia. [171]Department of Animal Health and Production, Oyo State College of Agriculture and Technology, Igbo-Ora 201103, Oyo State, Nigeria. [172]Department of Forest & Wildlife Ecology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI 53706, USA. [173]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, 00014 Helsinki, Finland. [174]South Iceland Research Centre, University of Iceland, Lindarbraut 4, 840 Laugarvatn, Iceland. [175]Department of Clinical Neuroscience, Karolinska Institutet, Nobels väg 9, 171 77 Stockholm, Sweden. [176]Animal Ecology and Tropical Biology, University of Würzburg, Biocenter-Am Hubland, 97074 Würzburg, Germany. [177]Hiram College, 11700 Dean St, Biology, Hiram, OH 44234, USA. [178]Department of Aquatic Resources, Swedish University of Agricultural Sciences, Almas allé 5, 756 51 Uppsala, Sweden. [179]Department of Earth Sciences, Montana State University, Bozeman, MT 59717, USA. [180]Department of Evolution, Ecology, and Organismal Biology, University of California, 900 University Ave, RiversideRiverside, CA 92521, USA. [181]Sección Ornitología, Universidad Nacional de La Plata, Paseo del Bosqur s/n, La Plata, B1900FWA Buenos Aires, Argentina. [182]Bangor University, Bangor University, Deiniol Road, Bangor LL57 2UW, UK. [183]MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Robinson Way, Cambridge CB2 0SR, UK. [184]Harvard Forest, Harvard University, 324 N Main St, Petersham, MA 01366, USA. [185]Departamento de Biodiversidad, Ecología y Evolución, Universidad Complutense de Madrid, C. de José Antonio Novais, 12, 28040 Madrid, Spain. [186]Biology Department, Technische Universität Darmstadt, Schnittspahnstraße 3, 64287 Darmstadt, Germany. [187]UMR 1309, ASTRE, CIRAD, Campus international de Baillarguet, 34398 Montpellier, France. [188]Department of Entomology, The Ohio State University, 1680 Madison Ave, Wooster, OH 44691, USA. [189]Department of Psychology, University of Bath, 10 West, Bath BA2 7AY, UK. [190]Chaire de recherche en intégrité écologique, Université du Québec à Trois-Rivières, 3351 Boul. Des Forges, Trois-Rivières, QC G8Z 4M3, Canada. [191]Mississippi Based RESTORE Act Center of Excellence, University of Southern Mississippi, 703 E. Beach Drive, Ocean Springs, MS 39564, USA. [192]Department of Integrative Biology, University of California, Valley Life Science Building 5075, BerkeleyBerkeley, CA 94720, USA. [193]Mention Zoologie et Biodiversité Animale, Faculté des Sciences, Université d'Antananarivo, Mention Zoologie et Biodiversié Animale, Université d'Antananarivo, BP 906, 101 Antananarivo, Madagascar. [194]Department of Integrative Biology, Valley Life Sciences Building 3140, University of California, University of California Berkeley, BerkeleyBerkeley, CA 94720, USA. [195]Department of Ecology, Behavior and Evolution, University of California, San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA. [196]Institute for Environmental Sciences, VU Amsterdam, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands. [197]Department of Evolutionary Anthropology, University of Vienna, Djerassiplatz 1 (UBB), 1030 Wien, Austria. [198]Konrad Lorenz Institute for Ethology, University of Veterinary Medicine, Savoyenstrasse 1A, 1160 Vienna, Austria. [199]School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China. [200]Institut de biologie, Université de Neuchâtel, Emile-Argand 11, 2000 Neuchâtel, Switzerland. [201]School of Information, University of Arizona, 1103 E. 2nd St, Tucson, AZ 85721, USA. [202]Center for Biological Data Science, Virginia Commonwealth University, 1000 W. Cary St, Box 842030, Richmond, VA 23284-2030, USA. [203]University of Wisconsin, 1525 Observatory Dr. Madison, Madison, WI 53706, USA. [204]School of the Environment, Yale University, 195 Prospect Street, New Haven, CT 06511, USA. [205]Institute of the Environment, Florida International University, 3000 NE 151st St, North Miami, FL 33181, USA. [206]Department of Biology, Missouri State University, 910 S John Q Hammons Pkwy, Springfield, MO 65897, USA. [207]Department of Biological and Environmental Science, University of Jyväskylä, Survontie 9C, 40500 Jyväskylä, Finland. [208]Institute for Interdisciplinary Mountain Research, OeAW (Austrian Academy of Sciences), GLORIA, Silbergasse 30/3, A-1190 Wien, Austria. [209]Department of Natural Resources, Newe Ya'ar Research Center, Agricultural Research Organization (Volcani Institute), POB 1021, 3009500 Ramat Yishay, Israel. [210]Department of Animal Behaviour, Bielefeld University, Konsequenz 45, 33615 Bielefeld, Germany. [211]Office for National Statistics, Segensworth Rd, Titchfield, Fareham PO15 5RR, UK. [212]Institute of Avian Research "Vogelwarte Helgoland", An der Vogelwarte 21, 26386 Wilhelmshaven, Germany. [213]Department of Evolutionary Biology, Bielefeld University, North Rhine-Westphalia, Konsequenz 45, 33615 Bielefeld, Germany. [214]Ecology Department, Universidade Federal de Goiás, Av. Esperança, Campus Samambaia, Goiânia, Goiás 74690-900, Brazil. [215]Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, Cornwall TR10 9FE, UK. [216]New South Wales, Department of Primary Industries Fisheries, Locked Bag 1, Nelson

Bay, NSW 2315, Australia. [217]Research Data Management, Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder Straße 84, 15374 Müncheberg, Germany. [218]Biology Department, University of Wisconsin-La Crosse, 1725 State St, La Crosse, WI 54601, USA. [219]Department of Evolutionary Anthropology, Duke University, 130 Science Dr, Durham, NC 27708, USA. [220]Earth and Life Institute, Ecology and Biodiversity, UCLouvain, Croix du Sud 4, L7.07.04, 1348 Louvain-la-Neuve, Belgium. [221]Future Regions Research Centre, Federation University Australia, Mt Helen, VIC 3350, Australia. [222]United States, Department of Agriculture- Agricultural Research Service, 1701 10th Ave SW, Mandan, ND 58554, USA. [223]Arthur Rylah Institute for Environmental Research, 123 Brown Street, Heidelberg, Victoria 3084, Australia. [224]Epidemiology and Surveillance Support Unit, University of Lyon - French Agency for Food, Environmental and Occupational Health and Safety (ANSES), 31 Avenue Tony Garnier, 69007 Lyon, France. [225]Center for Impact, UCLA Anderson, University of California, 110 Westwood Plaza, Gold Hall, Suite B.201L, Los AngelesLos Angeles, CA 90095-1481, USA. [226]Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Montevideo, Uruguay. [227]Washington, USA. [228]Programa de Pós-Graduação em Ecologia, Instituto de Biologia, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Cidade Universitária, Av. Carlos Chagas Filho 373, Rio de Janeiro, RJ, CEP 21941-902, Brazil. [229]Vive Crop Protection, 6275 Northam Drive, Suite 1, Mississauga, ON L4V 1Y8, Canada. [230]University of Cambridge, Trinity Ln, The Old Schools, Cambridge CB2 1TN, UK. [231]British Trust for Ornithology, BTO, The Nunnery, Thetford, Norfolk IP24 2PU, UK. [232]Technology & Society Department, Rochester Institute of Technology, 7 Lomb Memorial Drive, Rochester, NY 14623, USA. [233]Nomad Ecology, 822 Main Street, Martinez, CA 94553, USA. [234]Wildland Resources Department, Utah State University, 5200 Old Main Hill, Logan, UT 84322, USA. [235]Center for Biological Control, Department of Zoology and Entomology, Rhodes University, 1 Lower University Road, Barratt Complex, Biological Sciences BuildingEastern Cape, Makhanda, South Africa. [236]School of Mathematics and Statistics and Centre for Research in Ecological and Environmental Modelling, University of St Andrews, Buchanan Gardens, St Andrews, Scotland KY16 9LZ, UK. [237]Department of Ecology and Evolutionary Biology, Cornell University, 215 Tower Road, Ithaca, NY 14853, USA. [238]Department of Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Permoserstraße 15, 04318 Leipzig, Germany. [239]Ecosystems and Global Change Group, School of the Environment, Trent University, 1600 West Bank Road, Peterborough, Ontario K0L 2V0, Canada. [240]Instituto Universitario de Investigación en Gestión Forestal Sostenible (iuFOR), Universidad de Valladolid, Av. Madrid 44, 34071 Palencia, Spain. [241]Department of Biological Sciences, University of Bergen, Postboks, 7803, N-5020 Bergen, Norway. [242]Department of Biological Science, University of Rhode Island, 9 East Alumni Ave, Kingston, RI 02881, USA. [243]Department of Geography, McGill University, 805 Sherbrooke Street West, Montreal, Quebec H3A 0B9, Canada. [244]School of Biological and Marine Sciences, University of Plymouth, Drake Circus, Plymouth, Devon PL4 8AA, UK. [245]Biology Department, Wake Forest University, 1834 Wake Forest Rd., Winston Salem, NC 27109, USA. [246]Plant Sciences, University of California, 1 Shields Ave, DavisDavis, CA 95616, USA. [247]College of Natural Resources, North Carolina State University, Jordan Hall, 2800 Faucette Dr, Raleigh, NC 27607, USA. [248]Institute of Zoology, Technische Universität Dresden, Zellescher Weg 20b, 01217 Dresden, Germany. [249]Helmholtz AI, Helmholtz Zentrum Muenchen, Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany. [250]FitzPatrick Institute of African Ornithology, University of Cape Town, University of Cape Town, Private Bag X3, Rondebosch, Cape Town 7701, South Africa. [251]Department of Cell & Developmental Biology, Division of Biosciences, University College London, London, UK. [252]Biology, University of Saskatchewan, University of Saskatchewan, 112 Science Place, Saskatoon, SK S7N 5E2, Canada. [253]Department of Biology, University of Regina, 3737 Wascana Pkwy, Regina, Saskatchewan S4S 0A2, Canada. [254]Biology, University of Waterloo, 200 University Ave W, Waterloo, Ontario N2L 3G1, Canada. [255]Department of Ecology & Evolutionary Biology, Biological Science Building, University of Michigan, 1105 North University Avenue, Ann Arbor, MI 48109-1085, USA. [256]Dept. Ecologia, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Av. Carlos Chagas Filho 373, Rio de Janeiro/RJ, CP 6820021942-902, Brazil. [257]Lothian Analytical Services, Public Health Scotland, 1 South Gyle Crescent, Edinburgh EH12 9EB, UK. [258]Institute for Evolution and Biodiversity, University of Muenster, Huefferstr. 1, DE-48149 Muenster, Germany. [259]Department of Environmental Sciences, Western Norway University of Applied Sciences, P.O. box 133, 6851 Sogndal, Norway. [260]Department of Biological Sciences, University of Southern Maine, 70 Falmouth St, Portland, ME 04103, USA. [261]Center for Ecosystem Science and Society, Northern Arizona University, PO Box 5620, Flagstaff, AZ 86011, USA. [262]Center for Watershed Sciences, University of California, Davis, 1 Shields Ave, Davis, CA 95616, USA. [263]School of Agriculture and Environmental Science, University of Southern Queensland, 487-535 West Street, Toowoomba, Qld 4350, Australia. [264]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, 2200 Osborn Dr, Ames, IA 50011, USA. [265]Fram Project AS, Ymers vei 2, 0588 Oslo, Norway. [266]Department of Food Science & Technology, Virginia Polytechnic Institute and State University, 22 Food Science Building (0418) 360 Duck Pond Drive Virginia Tech, Blacksburg, VA 24061, USA. [267]School of Applied Sciences, School of Applied Sciences, University of Brighton, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK. [268]Department of Biology, Biology Research and Administration Building, University of Oxford, 11a Mansfield Rd, Oxford OX1 3SZ, UK. [269]Department of Integrative Biology, University of Colorado, P.O. Box 173364, DenverDenver, CO 80217-3364, USA. [270]Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, MI 48824, USA. [271]ISEM, University of Montpellier, CNRS, Place Eugène Bataillon-Cedex 05, 34095 Montpellier, France. [272]Laboratoire d'Ethologie Expérimentale et Comparée, LEEC, Université Sorbonne Paris Nord, 99 avenue Jean-Baptiste Clément, UR444393430 Villetaneuse, France. [273]Department of Science and Environment, Lake Superior State University, 650 W Easterday Ave, Sault Sainte Marie, MI 49783, USA.

Received: 17 August 2019 Accepted: 19 December 2024
Published online: 06 February 2025

References

1. Aczel, Balazs, Barnabas Szaszi, Gustav Nilsonne, Olmo R van den Akker, Casper J Albers, Marcel ALM van Assen, Jojanneke A Bastiaansen, et al. 2021. "Consensus-Based Guidance for Conducting and Reporting Multi-Analyst Studies. eLife. 10 (November). https://doi.org/10.7554/elife.72185.
2. Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2024. Quarto. https://doi.org/10.5281/zenodo.5960048.
3. Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. rmarkdown: Dynamic Documents for r. https://github.com/rstudio/rmarkdown.
4. Arif S, Aaron MacNeil M. Predictive models aren't for causal inference. Ecology Letters. 2022;25(8):1741–5. https://doi.org/10.1111/ele.14033.
5. Arif, Suchinta, and M. Aaron MacNeil. 2023. "Applying the Structural Causal Model Framework for Observational Causal Inference in Ecology." Ecological Monographs. 93(1):e1554. https://doi.org/10.1002/ecm.1554.
6. Arnold, Jeffrey B. 2024. ggthemes: Extra Themes, Scales and Geoms for "ggplot2". https://jrnold.github.io/ggthemes/.
7. Atkinson, Joe, Lars A. Brudvig, Max Mallen-Cooper, Shinichi Nakagawa, Angela T. Moles, and Stephen P. Bonser. 2022. "Terrestrial Ecosystem Restoration Increases Biodiversity and Reduces Its Variability, but Not to Reference Levels: A Global Meta-Analysis. Ecology Letters. 25 (7): 1725–37. https://doi.org/10.1111/ele.14025.
8. Auspurg Katrin, Brüderl Josef. Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the 'Many Analysts, One Data Set' Project. Socius. 2021;7:23780231211024420. https://doi.org/10.1177/23780231211024421.
9. Bartoń, Kamil. MuMIn: Multi-Model Inference. R package version 1.47.5. 2023. https://cran.r-project.org/web/packages/MuMIn.
10. Baselga, Andres, David Orme, Sebastien Villeger, Julien De Bortoli, Fabien Leprieur, Maxime Logez, Sara Martinez-Santalla, et al. 2023. betapart: Partitioning Beta Diversity into Turnover and Nestedness Components. https://CRAN.R-project.org/package=betapart.
11. Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4. 2015;67(1):48. https://doi.org/10.18637/jss.v067.i01.
12. Blake, Kevin. 2022. NatParksPalettes: Color Palettes Inspired by National Parks. https://github.com/kevinsblake/NatParksPalettes.
13. Bolker, Ben, David Robinson, Dieter Menne, Jonah Gabry, Paul Buerkner, Chrisopher Hau, William Petry, et al. 2024. broom.mixed: Tidying Methods for Mixed Models. https://github.com/bbolker/broom.mixed.

14. Borenstein Michael, Higgins Julian P. T, Hedges Larry, Rothstein Hannah. Basics of Meta-Analysis: I^2 Is Not an Absolute Measure of Heterogeneity. Research Synthesis Methods. 2017;8:5–18. https://doi.org/10.1002/jrsm.1230.
15. Botvinik-Nezer Rotem, Holzmeister Felix, Camerer Colin F, Dreber Anna, Huber Juergen, Johannesson Magnus, Kirchler Michael, et al. Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams. Nature. 2020;582(7810):84–8.
16. Breznau Nate, Rinke Eike Mark, Wuttke Alexander, Nguyen Hung H. V, Adem Muna, Adriaans Jule, Alvarez-Benjumea Amalia, et al. Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty. Proceedings of the National Academy of Sciences. 2022;119(44): e2203150119. https://doi.org/10.1073/pnas.2203150119.
17. Briga, Michael, and Simon Verhulst. 2021. "Mosaic Metabolic Ageing: Basal and Standard Metabolic Rates Age in Opposite Directions and Independent of Environmental Quality, Sex and Life Span in a Passerine. Functional Ecology. 35 (5): 1055–68. https://doi.org/10.1111/1365-2435.13785.
18. Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling. The R Journal. 9(2):378–400. https://doi.org/10.32614/RJ-2017-066.
19. Buck Robert J, Fieberg John, Larkin Daniel J. The use of weighted averages of Hedges' d in meta-analysis: Is it worth it? Methods in Ecology and Evolution. 2022;13(5):1093–105. https://doi.org/10.1111/2041-210X.13818.
20. Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. Book. 2nd ed. New York: Springer-Verlag. https://doi.org/10.1007/b97636.
21. Cade, Brian S. 2015. "Model Averaging and Muddled Multimodel Inferences. Ecology. 96(9):2370–82. http://www.jstor.org.ezproxy.whitman.edu/stable/24702343.
22. Capilla-Lasheras Pablo, Thompson Megan J, Sánchez-Tójar Alfredo, Haddou Yacob, Branston Claire J, Réale Denis, Charmantier Anne, Dominoni Davide M. A Global Meta-Analysis Reveals Higher Variation in Breeding Phenology in Urban Birds Than in Their Non-Urban Neighbours. Ecology Letters. 2022;25(11):2552–70. https://doi.org/10.1111/ele.14099.
23. Coretta Stefano, Casillas Joseph V, Roessig Simon, Franke Michael, Ahn Byron, Al-Hoorie Ali H, Al-Tamimi Jalal, et al. Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses. Advances in Methods and Practices in Psychological Science. 2023;6(3):25152459231162570. https://doi.org/10.1177/25152459231162567.
24. Dancho, Matt, and Davis Vaughan. 2023. Timetk: A Tool Kit for Working with Time Series. https://CRAN.R-project.org/package=timetk.
25. DeKogel, C. H. 1997. "Long-Term Effects of Brood Size Manipulation on Morphological Development and Sex-Specific Mortality of Offspring. Journal of Animal Ecology. 66(2):167–78. <Go to ISI>://WOS:A1997WQ19600003.
26. Deressa, Teshome, David Stern, Jaco Vangronsveld, Jan Minx, Sebastien Lizin, Robert Malina, and Stephan Bruns. 2023. "More Than Half of Statistically Significant Research Findings in the Environmental Sciences Are Actually Not. EcoEvoRxiv. https://doi.org/10.32942/X24G6Z.
27. Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, et al. 2013. "Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance. Ecography. 36(1):27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.
28. Fanelli Daniele, Costas Rodrigo, Ioannidis John P. A. Meta-Assessment of Bias in Science. Proceedings of the National Academy of Sciences. 2017;114:3714–9. https://doi.org/10.1073/pnas.1618569114.
29. Fanelli Daniele, Ioannidis John P. A. US Studies May Overestimate Effect Sizes in Softer Research. Proceedings of the National Academy of Sciences. 2013;110(37):15031–6. https://doi.org/10.1073/pnas.1302997110.
30. Fidler Fiona, Burgman Mark A, Cumming Geoff, Buttrose Robert, Thomason Neil. Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. Conservation Biology. 2006;20(5):1539–44. https://doi.org/10.1111/j.1523-1739.2006.00525.x.
31. Fidler Fiona, Chee Yung En, Wintle Bonnie C, Burgman Mark A, McCarthy Michael A, Gordon Ascelin. Metaresearch for Evaluating Reproducibility in Ecology and Evolution. BioScience. 2017;67(3):282–9. https://doi.org/10.1093/biosci/biw159.
32. Firke, Sam. 2023. janitor: Simple Tools for Examining and Cleaning Dirty Data. https://github.com/sfirke/janitor.
33. Forstmeier Wolfgang, Wagenmakers Eric-Jan, Parker TH. Detecting and Avoiding Likely False-Positive Findings – a Practical Guide. Biological Reviews. 2017;92:1941–68. https://doi.org/10.1111/brv.12315.
34. Fraser Hannah, Parker Tim, Nakagawa Shinichi, Barnett Ashley, Fidler Fiona. Questionable Research Practices in Ecology and Evolution. PLOS ONE. 2018;13(7): e0200303. https://doi.org/10.1371/journal.pone.0200303.
35. Gamer, Matthias, Jim Lemon, and Ian Fellows Puspendra Singh. 2019. irr: Various Coefficients of Interrater Reliability and Agreement. https://www.r-project.org.
36. Gelman Andrew, Loken Eric. The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-Hacking' and the Research Hypothesis Was Posited Ahead of Time. Department of Statistics: Columbia University; 2013.
37. Gelman Andrew, Weakliem David. Of Beauty, Sex, and Power. American Scientist. 2009;97:310–6.
38. Gould Elliot, Fraser Hannah S, Nakagawa Shinichi, Parker Timothy H. ManyEcoEvo: Meta-Analyse Data from ManyAnalyst Style Studies. 2023. Zenodo. https://doi.org/10.5281/zenodo.10046153.
39. Gould Ellliot, Fraser Hannah S, Nakagawa Shinichi, Parker Timothy H. egouldo/ManyAnalysts: Manuscript Source Code for "Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology." 2024. Zenodo. https://doi.org/10.5281/zenodo.13850927.Version2.0.2.
40. Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. "Multimodel Inference in Ecology and Evolution: Challenges and Solutions. Journal of Evolutionary Biology. 24(4):699–711. https://doi.org/10.1111/j.1420-9101.2010.02210.x.
41. Harrell Jr, Frank E. 2024. Hmisc: Harrell Miscellaneous. https://hbiostat.org/R/Hmisc/.
42. Hester, Jim, Lionel Henry, Kirill Müller, Kevin Ushey, Hadley Wickham, and Winston Chang. 2024. withr: Run Code "With" Temporarily Modified Global State. https://withr.r-lib.org.
43. Higgins Julian P T, Thompson Simon G, Deeks Jonathan J, Altman Douglas G. Measuring Inconsistency in Meta-Analyses. BMJ. 2003;327(7414):557–60. https://doi.org/10.1136/bmj.327.7414.557.
44. Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics. Economic Inquiry. 59(3):944–60. https://doi.org/10.1111/ecin.12992.
45. Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2024. gt: Easily Create Presentation-Ready Display Tables. https://gt.rstudio.com.
46. Jennions, M. D., C. J. Lortie, M. S. Rosenberg, and H. R. Rothstein. 2013. "Publication and Related Biases." Book Section. In Handbook of Meta-Analysis in Ecology and Evolution, edited by J. Koricheva, J. Gurevitch, and K. Mengersen, 207–36. Princeton, USA: Princeton University Press.
47. Kassambara, Alboukadel. 2023. ggpubr: "ggplot2" Based Publication Ready Plots. https://rpkgs.datanovia.com/ggpubr/.
48. Kimmel Kaitlin, Avolio Meghan L, Ferraro Paul J. Empirical Evidence of Widespread Exaggeration Bias and Selective Reporting in Ecology. Nature Ecology & Evolution. 2023. https://doi.org/10.1038/s41559-023-02144-3.
49. Klein Richard A, Ratliff Kate A, Vianello Michelangelo, Adams Jr Reginald B, Bahník Štěpán, Bernstein Michael J, Bocian Konrad, et al. Investigating Variation in Replicability: A "Many Labs" Replication Project. Social Psychology. 2014;45(3):142–52. https://doi.org/10.1027/1864-9335/a000178.
50. Klein Richard A, Vianello Michelangelo, Hasselman Fred, Adams Byron G, Adams Reginald B, Alper Sinan, Aveyard Mark, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and

Settings. Advances in Methods and Practices in Psychological Science. 2018;1(4):443–90. https://doi.org/10.1177/2515245918810225.
51. Knight K. *Mathematical Statistics*. Book. New York: Chapman; Hall; 2000.
52. Koricheva, Julia, and Jessica Gurevitch. 2014. "Uses and Misuses of Meta-Analysis in Plant Ecology. Journal of Ecology. 102(4):828–44. https://doi.org/10.1111/1365-2745.12224.
53. Kou-Giesbrecht, Sian, and Duncan N. L. Menge. 2021. "Nitrogen-Fixing Trees Increase Soil Nitrous Oxide Emissions: A Meta-Analysis. Ecology. 102(8):e03415. https://doi.org/10.1002/ecy.3415.
54. Kuhn, Max, and Hannah Frick. 2022. multilevelmod: Model Wrappers for Multi-Level Models. https://github.com/tidymodels/multilevelmod.
55. Kuhn, Max, and Hadley Wickham. 2020. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles. https://www.tidymodels.org.
56. Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software. 82(13):1–26. https://doi.org/10.18637/jss.v082.i13.
57. Landau, William Michael. 2021. "The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing. Journal of Open Source Software. 6(57):2959. https://doi.org/10.21105/joss.02959.
58. Leybourne, Daniel J., Katharine F. Preedy, Tracy A. Valentine, Jorunn I. B. Bos, and Alison J. Karley. 2021. "Drought Has Negative Consequences on Aphid Fitness and Plant Vigor: Insights from a Meta-Analysis. Ecology and Evolution. 11(17):11915–29. https://doi.org/10.1002/ece3.7957.
59. Lu, Xun, and Halbert White. 2014. "Robustness Checks and Robustness Tests in Applied Economics. Journal of Econometrics. 178:194–206. https://doi.org/10.1016/j.jeconom.2013.08.016.
60. Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. "Extracting, Computing and Exploring the Parameters of Statistical Models Using R. Journal of Open Source Software. 5(53):2445. https://doi.org/10.21105/joss.02445.
61. Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "performance: An R Package for Assessment, Comparison and Testing of Statistical Models. Journal of Open Source Software. 6(60):3139. https://doi.org/10.21105/joss.03139.
62. Lüdecke, Daniel, Indrajeet Patil, Mattan S. Ben-Shachar, Brenton M. Wiernik, Philip Waggoner, and Dominique Makowski. 2021. "see: An R Package for Visualizing Statistical Models. Journal of Open Source Software. 6(64):3393. https://doi.org/10.21105/joss.03393.
63. Luke SG. Evaluating Significance in Linear Mixed-Effects Models in r. Behavior Research Methods. 2017;49(4):1494–502.
64. Makowski, Dominique, Mattan S. Ben-Shachar, Indrajeet Patil, and Daniel Lüdecke. 2020. "Estimation of Model-Based Predictions, Contrasts and Means. CRAN. https://github.com/easystats/modelbased.
65. Masur, Philipp K., and Michael Scharkow. 2020. "specr: Conducting and Visualizing Specification Curve Analyses (Version 1.0.0)." https://CRAN.R-project.org/package=specr.
66. Meschiari, Stefano. 2022. Latex2exp: Use LaTeX Expressions in Plots. https://www.stefanom.io/latex2exp/.
67. Miles, C. 2008. "Testing Market-Based Instruments for Conservation in Northern Victoria." Book Section. In Biodiversity: Integrating Conservation and Production: Case Studies from Australian Farms, Forests and Fisheries, edited by T. Norton, T. Lefroy, K. Bailey, and G. Unwin, 133–46. Melbourne, Australia: CSIRO Publishing.
68. Millard, Steven P. 2013. EnvStats: An r Package for Environmental Statistics. New York: Springer. https://www.springer.com.
69. Molina, Isabel, and Yolanda Marhuenda. 2015. "sae: An R Package for Small Area Estimation. The R Journal. 7(1):81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf.
70. Morrissey, Michael B., and Graeme D. Ruxton. 2018. "Multiple Regression Is Not Multiple Regressions: The Meaning of Multiple Regression and the Non-Problem of Collinearity. Philosophy, Theory, and Practice in Biology. 10(3). https://doi.org/10.3998/ptpbio.16039257.0010.003.
71. Müller, Kirill. 2020. here: A Simpler Way to Find Your Files. https://here.r-lib.org/.
72. Nakagawa Shinichi, Cuthill Innes C. Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists. Biological Reviews. 2007;82(4):591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x.
73. Nakagawa, Shinichi, Malgorzata Lagisz, Michael D. Jennions, Julia Koricheva, Daniel W. A. Noble, Timothy H. Parker, Alfredo Sánchez-Tójar, Yefeng Yang, and Rose E. O'Dea. 2022. "Methods for Testing Publication Bias in Ecological and Evolutionary Meta-Analyses. Methods in Ecology and Evolution. 13(1):4–21. https://doi.org/10.1111/2041-210X.13724.
74. Nakagawa, Shinichi, Malgorzata Lagisz, Rose E. O'Dea, Patrice Pottier, Joanna Rutkowska, Alistair M. Senior, Yefeng Yang, and Daniel W. A. Noble. 2023. "orchaRd 2.0: An r Package for Visualizing Meta-Analyses with Orchard Plots. EcoEvoRxiv. 12:4–12. https://doi.org/10.32942/X2QC7K.
75. Nakagawa Shinichi, Yang Yefeng, Macartney Erin L, Spake Rebecca, Lagisz Malgorzata. Quantitative Evidence Synthesis: A Practical Guide on Meta-Analysis, Meta-Regression, and Publication Bias Tests for Environmental Sciences. Environmental Evidence. 2023;12(1):8. https://doi.org/10.1186/s13750-023-00301-6.
76. Nakagawa S, Noble DW, Senior AM, Lagisz M. Meta-Evaluation of Meta-Analysis: Ten Appraisal Questions for Biologists. BMC Biology. 2017;15(1):18. https://doi.org/10.1186/s12915-017-0357-7.
77. Nicolaus M, Michler SPM, Ubels R, van der Velde M, Komdeur J, Both C, Tinbergen JM. Sex-Specific Effects of Altered Competition on Nestling Growth and Survival: An Experimental Manipulation of Brood Size and Sex Ratio. Journal of Animal Ecology. 2009;78(2):414–26. https://doi.org/10.1111/j.1365-2656.2008.01505.x.
78. Noble Daniel W. A, Lagisz Malgorzata, O'Dea Rose E, Nakagawa Shinichi. Nonindependence and Sensitivity Analyses in Ecological and Evolutionary Meta-Analyses. Molecular Ecology. 2017;26(9):2410–25. https://doi.org/10.1111/mec.14031.
79. O'Hara Robert B, Johan Kotze D. Do Not Log-Transform Count Data. Methods in Ecology and Evolution. 2010;1(2):118–22. https://doi.org/10.1111/j.2041-210x.2010.00021.x.
80. Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." Science. 349(6251):aac4716. https://doi.org/10.1126/science.aac4716.
81. Page Matthew J, Moher David. Evaluations of the Uptake and Impact of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement and Extensions: A Scoping Review. Systematic Reviews. 2017;6(1):263. https://doi.org/10.1186/s13643-017-0663-8.
82. Parker Timothy H, Forstmeier Wolfgang, Koricheva Julia, Fidler Fiona, Hadfield Jarrod D, Chee Yung En, Kelly Clint D, Gurevitch Jessica, Nakagawa Shinichi. Transparency in Ecology and Evolution: Real Problems, Real Solutions. Trends in Ecology & Evolution. 2016;31(9):711–9. https://doi.org/10.1016/j.tree.2016.07.002.
83. Parker Timothy H, Yang Yefeng. Exaggerated Effects in Ecology. Nature Ecology & Evolution. 2023. https://doi.org/10.1038/s41559-023-02156-z.
84. Pedersen, Thomas Lin. 2024. patchwork: The Composer of Plots. https://patchwork.data-imaginist.com.
85. Pei Yifan, Forstmeier Wolfgang, Wang Daiping, Martin Katrin, Rutkowska Joanna, Kempenaers Bart. Proximate Causes of Infertility and Embryo Mortality in Captive Zebra Finches. The American Naturalist. 2020;196(5):577–96. https://doi.org/10.1086/710956.
86. Qiu, Yixuan. 2024. showtext: Using Fonts More Easily in r Graphs. https://github.com/yixuan/showtext.
87. R Core Team. 2024. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
88. Rosenberg, M. S. 2013. "Moment and Least-Squares Based Approaches to Metaanalytic Inference." Book Section. In Handbook of Meta-Analysis in Ecology and Evolution, edited by J. Koricheva, J. Gurevitch, and K. Mengersen, 108–24. Princeton, USA: Princeton University Press.
89. Royle NJ, Hartley IR, Owens IPF, Parker GA. Sibling Competition and the Evolution of Growth Rates in Birds. Proceedings of the Royal Society B-Biological Sciences. 1999;266(1422):923–32. https://doi.org/10.1098/rspb.1999.0725.
90. Scheinin, Ilari, Maria Kalimeri, Vilma Jagerroos, Juuso Parkkinen, Emmi Tikkanen, Peter Würtz, and Antti Kangas. 2020. ggforestplot: Forestplots of Measures of Effects and Their Confidence Intervals. https://github.com/NightingaleHealth/ggforestplot.
91. Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. GGally: Extension to "ggplot2". https://ggobi.github.io/ggally/.

92. Schweinsberg M, Feldman M, Staub N, van den Akker OR, van Aert RCM, Malm van Assen Y, Liu, et al. Same Data, Different Conclusions: Radical Dispersion in Empirical Results When Independent Analysts Operationalize and Test the Same Hypothesis. Organizational Behavior and Human Decision Processes. 2021;165:228–49. https://doi.org/10.1016/j.obhdp.2021.02.003.
93. Senior Alistair M, Grueber Catherine E, Kamiya Tsukushi, Lagisz Malgorzata, O'Dwyer Katie, Santos Eduardo S. A, Nakagawa Shinichi. Heterogeneity in Ecological and Evolutionary Meta-Analyses: Its Magnitude and Implications. Ecology. 2016;97(12):3293–9. https://doi.org/10.1002/ecy.1591.
94. Shavit, A., and Aaron M. Ellison. 2017. Stepping in the Same River Twice: Replication in Biological Research. Edited Book. New Haven, Connecticut, USA: Yale University Press.
95. Siegel, Kyle R., Muskanjot Kaur, A. Calvin Grigal, Rebecca A. Metzler, and Gary H. Dickinson. 2022. "Meta-Analysis Suggests Negative, but pCO2-Specific, Effects of Ocean Acidification on the Structural and Functional Properties of Crustacean Biomaterials." Ecology and Evolution. 12(6):e8922. https://doi.org/10.1002/ece3.8922.
96. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, Bahník Š, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science. 2018;1(3):337–56. https://doi.org/10.1177/2515245917747646.
97. Silge, Julia, and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in r." JOSS. 1(3). https://doi.org/10.21105/joss.00037.
98. Simons Daniel J, Shoda Yuichi, Stephen Lindsay D. Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. Perspectives on Psychological Science. 2017. https://doi.org/10.1177/1745691617708863.
99. Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2015. "Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications." Manuscript. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2694998.
100. Simonsohn Uri, Simmons Joseph P, Nelson Leif D. Specification Curve Analysis. Nature Human Behaviour. 2020;4(11):1208–14. https://doi.org/10.1038/s41562-020-0912-z.
101. Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. "Reproducible Summary Tables with the Gtsummary Package." The R Journal. 13:570–80. https://doi.org/10.32614/RJ-2021-053.
102. Slowikowski, Kamil. 2024. ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2". https://ggrepel.slowkow.com/.
103. Stanton-Geddes, John, Cintia Gomes de Freitas, and Cristian de Sales Dambros. 2014. "In Defense of p Values: Comment on the Statistical Methods Actually Used by Ecologists." Ecology. 95(3):637–42. https://doi.org/10.1890/13-1156.1.
104. Steegen Sara, Tuerlinckx Francis, Gelman Andrew, Vanpaemel Wolf. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science. 2016;11(5):702–12. https://doi.org/10.1177/1745691616658637.
105. Taylor, James W., and Kathryn S. Taylor. 2023. "Combining Probabilistic Forecasts of COVID-19 Mortality in the United States." European Journal of Operational Research. 304(1):25–41. https://doi.org/10.1016/j.ejor.2021.06.044.
106. Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." Journal of Statistical Software. 105(7):1–31. https://doi.org/10.18637/jss.v105.i07.
107. Touchon, Justin C., and Michael W. McCoy. 2016. "The Mismatch Between Current Statistical Practice and Doctoral Training in Ecology." Ecosphere. 7(8):e01394. https://doi.org/10.1002/ecs2.1394.
108. Ushey, Kevin, and Hadley Wickham. 2023. renv: Project Environments. https://rstudio.github.io/renv/.
109. van den Brand, Teun. 2024. Ggh4x: Hacks for "ggplot2". https://github.com/teunbrand/ggh4x.
110. Werf Vander, Eric. Lack's Clutch Size Hypothesis: An Examination of the Evidence Using Meta-Analysis. Ecology. 1992;73(5):1699–705. https://doi.org/10.2307/1940021.
111. Hoef Ver, Jay M. Who Invented the Delta Method? The American Statistician. 2012;66(2):124–7. https://doi.org/10.1080/00031305.2012.687494.
112. Verhulst S, Holveck MJ, Riebel K. Long-Term Effects of Manipulated Natal Brood Size on Metabolic Rate in Zebra Finches. Biology Letters. 2006;2(3):478–80. https://doi.org/10.1098/rsbl.2006.0496.
113. Vesk PA, Morris WK, McCallum W, Apted R, Miles C. Processes of Woodland Eucalypt Regeneration: Lessons from the Bush Returns Trial. Proceedings of the Royal Society of Victoria. 2016;128:54–63.
114. Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the metafor Package." Journal of Statistical Software. 36(3):1–48. https://doi.org/10.18637/jss.v036.i03.
115. Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software. 4(43):1686. https://doi.org/10.21105/joss.01686.
116. Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. devtools: Tools to Make Developing r Packages Easier. https://devtools.r-lib.org/.
117. Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. scales: Scale Functions for Visualization. https://scales.r-lib.org.
118. Wilke, Claus O. 2024. cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". https://wilkelab.org/cowplot/.
119. Xie, Yihui. 2024a. knitr: A General-Purpose Package for Dynamic Report Generation in r. https://yihui.org/knitr/.
120. ———. 2024b. xfun: Supporting Functions for Packages Maintained by "Yihui Xie". https://github.com/yihui/xfun.
121. Yang Yefeng, Sánchez-Tójar Alfredo, O'Dea Rose E, Noble Daniel W. A, Koricheva Julia, Jennions Michael D, Parker Timothy H, Lagisz Malgorzata, Nakagawa Shinichi. Publication Bias Impacts on Effect Size, Statistical Power, and Magnitude (Type m) and Sign (Type s) Errors in Ecology and Evolutionary Biology. BMC Biology. 2023;21(1):71. https://doi.org/10.1186/s12915-022-01485-y.
122. Zeileis, Achim, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. 2020. "colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes." Journal of Statistical Software. 96(1):1–49. https://doi.org/10.18637/jss.v096.i01.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

C Supporting Information for the "Questionable Research Practices in Ecological Modelling" Chapter

Supporting Information for Chapter 4.

C.1 QRPs Identified from Literature Review

Table C.1: *Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1.*

QRP Class	QRP ID	QRP description	Model phase	Model subphase
Model Fishing	5	Changing the model specification or otherwise continuing to optimise the model after already validating or evaluating it on the test or holdout data.	Model Construction Model Evaluation Model Application	Model Specification Model Performance Metric Model Tuning Outcome Variable
	9	Conducting multiple different analyses or model variations after observing model checking / model performance results, selectively reporting only those analyses that yield favourable results without disclosing the full range of analyses performed.	Model Evaluation	—
	13	Dredging for models in unconstrained model space, where model space is not informed by theory or consists of models that are biologically implausible.	Model Construction	Model Specification Model Selection
	22	Focusing only on data / models that seemingly supports expectation or hypotheses, and disregarding evidence that does not corroborate hypotheses or expectation (even if present).	Model Construction	—
	31	Over-simplifying models due to ideological stance rather than based on modelling objectives or performance measures linked to those objectives.	Model Construction	Model Specification
	35	Re-partitioning data after observing model evaluation or model checking results.	Model Evaluation	Data Selection Data Partitioning
	41	Trying out different outcome variables or model evaluation metrics unrelated to model objectives and selecting based on performance after fitting the model and/or observing model checking / model evaluation results.	Model Construction Model Evaluation Model Application	Modelling Approach Model Evaluation Approach Model Performance metric Model Performance Metric Outcome Variable
Overhyping	7	Claiming the model has greater generalisability or credibility than it does.	Model Application	Inference
	26	Misreporting correlative claims using causal language.	Model Application	Inference
	39	Selectively reporting comparisons that support a foregone conclusion.	Model Application	Model Results
Poor Practice	10	Constructing new model / using new modelling approach rather than applying pre-existing one that might be superior.	Model Construction	—
	15	Failing to define model prediction properties.	Model Evaluation	Model Performance Metric
	16	Failing to define or inadequately defining model purpose, framing and or scope.	Model Construction	—
	17	Failure to clearly define research question or give precise definition of parameter of interest.	Model Construction	Model Purpose Outcome Variable

Table C.1: *Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1.*

QRP Class	QRP ID	QRP description	Model phase	Model subphase
	18	Failure to establish relative weighting of performance measures prior to beginning modelling.	Model Evaluation	Performance Measure Weighting
	19	Failure to explicitly state the model purpose, and / or failure to establish a priori performance metrics and measures after establishing the model purpose before beginning modelling.	Model Evaluation Problem Formulation	Model Performance Metric Model Purpose
	20	Failure to use biologically informed / justified predictor variables.	Model Construction	Model Specification
	21	Failure to use theory in guiding model specification / using default controls in regression model that are uninformed by theory.	Model Construction	Model Specification Model Tuning
	24	For studies developing new methods or approaches, optimising method / approach to improve the performance against baseline methods.	Model Construction	Method Selection / Model Selection
	25	Improper use of model evaluation metrics (e.g. using an evaluation metric ill-suited to the stated model purpose).	Model Evaluation	Model Performance Metric
	28	Misspecification of random effects structure by premature pruning of random effects / choosing random effect structure based on data rather than study design.	Model Construction	Model Specification Model Selection
	29	Model Selection Bias: deciding post hoc which distributional assumptions should be accepted, i.e. performing preliminary tests for distributional assumptions on the same data used for model selection. For example, checking for zero-inflation or overdispersion on the same data used for model selection.	Model Construction	Model Selection
	32	Overfitting a model to calibration dataset by including too many moderators or predictors relative to the size and complexity of the dataset.	Model Construction	Model Specification Model Tuning
	33	Overfitting model to calibration data by adding additional parameters after observing fitted model.	Model Construction	Model Specification Model Tuning
	34	Overuse of inferior, familiar methods / failing to adopt new, superior or best-practice methods.	Model Construction	Modelling Approach
	36	Resubstitution: assessing model performance on training set only or failing to evaluate model on independent data, or on partitioned holdout data.	Model Evaluation Model Construction	Data Partitioning Model Tuning NA
	38	Selective sampling / biased sampling, e.g. convenience or opportunistic sampling.	Model Construction	Data collection
	42	Using bad or easily obtainable, or inappropriate data to develop model.	Model Construction	Data Processing
	43	Using information-theoretic approaches to perform in-sample model assessment.	Model Evaluation	Model Performance Metric
S-hacking	1	Bayes Factor hacking: optimising the model to obtain a Bayes Factor above the required threshold.	Model Evaluation Model Construction	Data Processing Model Performance Metric Model Specification

Table C.1: *Questionable research practices (QRPs) may occur in different phases and sub-phases of the ecological modelling process. QRPs were identified through literature review and classified into broader classes. For each unique practice (QRP ID), literature sources are detailed online at doi.org/10.26188/30773831.v1.*

QRP Class	QRP ID	QRP description	Model phase	Model subphase
	2	Changing model output format or transformation after observing results.	Model Application	Model Results
	3	Changing model output or evaluation metric thresholds after observing outcome.	Model Evaluation Model Application	Model Performance Metric Model Output
	4	Changing random seed and refitting model after seeing results to improve model performance.	Model Construction Model Evaluation	Model Specification Model Tuning
	6	Changing the relative weighting of model performance metrics after observing model results, when there are multiple model evaluation analyses.	Model Evaluation	Model Performance Metric
	11	Discretising continuous variables after observing model checking / model performance results.	Model Construction Model Evaluation Model Application	Data Processing
	12	Dredging fitted models for statistical significance or other outcome variable.	Model Construction Model Evaluation	Model Selection Model Performance Metric
Sample Curation	8	Collecting new data and refitting model after observing model evaluation / model checking results (optional stopping rules).	Model Construction Model Evaluation	Data Collection Data Processing
	23	Focusing only on data that seemingly supports expectation or hypotheses, and disregarding data that does not corroborate hypotheses or expectation (even if present).	Model Construction	—
	27	Missing data hacking: changing the strategy to handle missing data after fitting the model and observing model checking / model performance evaluation results.	Model Construction Model Evaluation	Data Processing
	30	Modifying exclusion criteria or excluding data points, such as outliers or other values, without justification and prior planning, i.e. after fitting model and observing model evaluation / model checking results.	Model Construction Model Evaluation	Data Processing Model Performance Metric
Selective Debugging	14	Error checking only when unexpected or anomalous results are produced.	Model Construction Model Evaluation Model Application	Model Verification
Selective Reporting	37	Selective reporting of robustness checks in support of main results.	Model Evaluation	Robustness Checks
	40	Selectively reporting performance metrics that increase perception of performance after fitting model and/or observing model evaluation / model checking results.	Model Construction Model Evaluation	Model Performance Metric

C.2 Synthetic Example Code

Listing C.1: Code used to generate synthetic worked example in Figure 4.3.

```
library(tidyverse)
library(easystats)
library(patchwork)
library(gt)
library(gtExtras)
library(grDevices)
library(marquee)

# ---- Simulate Data ----

# Generate synthetic data
set.seed(123)
n <- 100
habitat_quality <- runif(n, 0, 10)
# True relationship with some noise
abundance <- 2 + 1.5 * habitat_quality + rnorm(n, 0, 2)

data <- tibble(
  habitat_quality = habitat_quality,
  abundance = abundance
)

# Define management scenarios & expected values under each scenario
management_scenarios <- tibble(
  action = c("Action A", "Action B"),
  habitat_quality_mean = c(6.0, 6.5), # Small difference initially
  habitat_quality_sd = c(0.5, 0.5)
)

habitat_values <- management_scenarios |>
  rowwise() |>
  mutate(
    habitat_values = list(rnorm(1000, habitat_quality_mean, habitat_quality_sd))
  )

# Stage 1: Initial simple model (defensible)
```

```
model_initial <- lm(abundance ~ habitat_quality, data = data)

# Predict for management scenarios
pred_initial <- habitat_values |>
  rowwise() |>
  mutate(
    predictions = list(
      predict(
        model_initial,
        newdata = tibble(habitat_quality = habitat_values),
        interval = "prediction"
      )
    ),
    pred_mean = mean(predictions[, 1]),
    pred_lower = quantile(predictions[, 1], 0.025),
    pred_upper = quantile(predictions[, 1], 0.975),
    predictions = list(predictions[, 1]),
    model = "Initial Model",
    stage = "Stage 1: A Priori Model"
  )

# Stage 2: Overfitted model (model fishing)
# Add polynomial and interaction terms to artificially reduce uncertainty
model_overfitted <- lm(
  abundance ~ poly(habitat_quality, 3) +
    I(habitat_quality^2 * (habitat_quality > 5)),
  data = data
)

pred_overfitted <- habitat_values |>
  rowwise() |>
  mutate(
    predictions = list(
      predict(
        model_overfitted,
        newdata = tibble(habitat_quality = habitat_values),
        interval = "prediction"
      )
    ),
    pred_mean = mean(predictions[, 1]),
    pred_lower = quantile(predictions[, 1], 0.025),
```

```
    pred_upper = quantile(predictions[, 1], 0.975),
    stage = "Stage 2: Model Fishing",
    model = "Overfitted Model",
    predictions = list(predictions[, 1])
  )

# Stage 3: Scenario hacking - artificially increase difference
management_scenarios_hacked <- management_scenarios |>
  mutate(
    habitat_quality_mean = case_when(
      action == "Action A" ~ 5.5, # Artificially reduced
      action == "Action B" ~ 7.5 # Artificially increased
    )
  )

pred_hacked <- management_scenarios_hacked |>
  rowwise() |>
  mutate(
    habitat_values = list(rnorm(
      1000,
      habitat_quality_mean,
      habitat_quality_sd
    )),
    predictions = list(
      predict(
        model_initial,
        newdata = tibble(habitat_quality = habitat_values),
        interval = "prediction"
      )
    ),
    pred_mean = mean(predictions[, 1]),
    pred_lower = quantile(predictions[, 1], 0.025),
    pred_upper = quantile(predictions[, 1], 0.975),
    stage = "Stage 3: Scenario Hacking",
    model = "Scenario Hacked",
    predictions = list(predictions[, 1])
  )


# Get descriptive statistics for violin plots
all_predictions <- bind_rows(
```

```
  pred_hacked |>
    select(model, action, pred_mean, pred_lower, pred_upper, stage) |>
    mutate(
      action_color = case_when(
        action == "Action A" ~ "#0072B2",
        action == "Action B" ~ "#2C5F41"
      )
    ),
  pred_initial |>
    select(model, action, pred_mean, pred_lower, pred_upper, stage) |>
    mutate(
      action_color = case_when(
        action == "Action A" ~ "#56B4E9",
        action == "Action B" ~ "#009E73"
      )
    ),
  pred_overfitted |>
    select(model, action, pred_mean, pred_lower, pred_upper, stage) |>
    mutate(
      action_color = case_when(
        action == "Action A" ~ "#56B4E9",
        action == "Action B" ~ "#009E73"
      )
    )
) |>
  mutate(
    stage = factor(
      stage,
      levels = c(
        "Stage 1: A Priori Model",
        "Stage 2: Model Fishing",
        "Stage 3: Scenario Hacking"
      )
    )
  )

# Plot Coefficients
pred_distributions <- bind_rows(
  # Stage 1: Initial model
  pred_initial |>
    select(action, predictions, stage) |>
```

```
    unnest(predictions),
  # Stage 2: Overfitted model
  pred_overfitted |>
    rowwise() |>
    select(action, predictions, stage) |>
    unnest(predictions),
  # Stage 3: Scenario hacked
  pred_hacked |>
    select(action, predictions, stage) |>
    unnest(predictions)
) |>
  mutate(
    stage = factor(
      stage,
      levels = c(
        "Stage 1: A Priori Model",
        "Stage 2: Model Fishing",
        "Stage 3: Scenario Hacking"
      )
    ),
    action_color = case_when(
      stage %in%
        c("Stage 1: A Priori Model", "Stage 2: Model Fishing") &
        action == "Action A" ~ "#56B4E9",
      stage %in%
        c("Stage 1: A Priori Model", "Stage 2: Model Fishing") &
        action == "Action B" ~ "#009E73",
      stage == "Stage 3: Scenario Hacking" & action == "Action A" ~ "#0072B2",
      stage == "Stage 3: Scenario Hacking" & action == "Action B" ~ "#2C5F41"
    )
  )

# ---- Construct Plots ----
# Violin Plots
p1 <- ggplot(pred_distributions, aes(x = action, y = predictions)) +
  geom_violin(aes(fill = I(action_color)), alpha = 0.7, trim = FALSE) +
  geom_boxplot(aes(color = I(action_color)), width = 0.1, alpha = 0.8) +
  stat_summary(
    aes(color = I(action_color)),
    fun = mean,
    geom = "point",
```

```
    size = 3,
    shape = 18
  ) +
  facet_wrap(~stage, ncol = 3) +
  labs(
    y = "Predicted Species Abundance",
    x = "Management Action"
  ) +
  theme_minimal() +
  hrbrthemes::theme_ipsum_rc() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 12),
    axis.text.x = element_text(size = 14, angle = 45, hjust = 1),
    axis.title.x = element_text(size = 16),
    axis.title.y = element_text(size = 16),
    strip.text = element_text(size = 14),
    legend.position = "none"
  )



effect_sizes <- all_predictions |> # Calculate effect sizes at each stage
  select(stage, action, pred_mean) |>
  pivot_wider(names_from = action, values_from = pred_mean) |>
  mutate(
    difference = `Action B` - `Action A`,
    effect_size = difference / 2 # Rough standardization
  )

# Model comparison plot showing overfitting with management actions
model_comparison <- tibble(
  habitat_quality = seq(0, 10, 0.1)
) |>
  mutate(
    initial_pred = predict(model_initial, newdata = .),
    overfitted_pred = predict(model_overfitted, newdata = .)
  ) |>
  pivot_longer(
    cols = c(initial_pred, overfitted_pred),
    names_to = "model_type",
    values_to = "prediction"
```

```
  ) |>
  mutate(
    model_type = case_when(
      model_type == "initial_pred" ~ "Initial Model",
      model_type == "overfitted_pred" ~ "Overfitted Model"
    )
  )

p3 <- ggplot() +
  geom_point(
    data = data,
    aes(x = habitat_quality, y = abundance),
    alpha = 0.6,
    color = "grey50"
  ) +
  geom_line(
    data = model_comparison,
    aes(
      x = habitat_quality,
      y = prediction,
      color = model_type,
      linetype = model_type
    ),
    linewidth = 1
  ) +
  # Initial management actions
  geom_vline(
    xintercept = management_scenarios |>
      pluck("habitat_quality_mean", 1),
    linetype = "solid",
    color = "#56B4E9",
    linewidth = 1,
    alpha = 0.7
  ) +
  geom_vline(
    xintercept = management_scenarios |>
      pluck("habitat_quality_mean", 2),
    linetype = "solid",
    color = "#009E73",
    linewidth = 1,
    alpha = 0.7
```

```
  ) +
  # Add hacked actions
  geom_vline(
    xintercept = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 1),
    linetype = "dashed",
    color = "#0072B2",
    linewidth = 1.2
  ) +
  geom_vline(
    xintercept = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 2),
    linetype = "dashed",
    color = "#2C5F41",
    linewidth = 1.2
  ) +
  # Arrows showing the manipulation
  annotate(
    "segment",
    x = management_scenarios |>
      pluck("habitat_quality_mean", 1),
    xend = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 1),
    y = 16,
    yend = 16,
    arrow = arrow(length = unit(0.3, "cm")),
    color = "#0072B2",
    linewidth = 1
  ) +
  annotate(
    "segment",
    x = management_scenarios |>
      pluck("habitat_quality_mean", 2),
    xend = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 2),
    y = 16,
    yend = 16,
    arrow = arrow(length = unit(0.3, "cm")),
    color = "#2C5F41",
    linewidth = 1
  ) +
```

```
  labs(
    x = "Habitat Quality",
    y = "Predicted Species Abundance",
  ) +
  theme_minimal() +
  hrbrthemes::theme_ipsum_rc() +
  theme(
    legend.position = c(0.99, 0.01),
    legend.justification = c(1, 0),
    legend.background = element_rect(
      fill = "white",
      color = "black",
      linewidth = 0.5
    ),
    legend.margin = margin(5, 5, 5, 5),
    legend.text = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold"),
    plot.subtitle = element_text(size = 12),
    axis.text.x = element_text(size = 14, angle = 45, hjust = 1),
    axis.title.x = element_text(size = 16),
    axis.title.y = element_text(size = 16),
  ) +
  scale_color_manual(
    "Model Version",
    values = c("Initial Model" = "#E69F00", "Overfitted Model" = "#D55E00")
  ) +
  scale_linetype_manual(
    "Model Version",
    values = c("Initial Model" = "solid", "Overfitted Model" = "dashed")
  ) +
  # Action labels
  annotate(
    "text",
    x = management_scenarios |>
      pluck("habitat_quality_mean", 1),
    y = 19,
    label = "Initial\nAction A",
    color = "#56B4E9",
    size = 3.5,
    hjust = 1.1,
    fontface = "bold"
```

```
  ) +
  annotate(
    "text",
    x = management_scenarios |>
      pluck("habitat_quality_mean", 2),
    y = 19,
    label = "Initial\nAction B",
    color = "#009E73",
    size = 3.5,
    hjust = -0.3,
    fontface = "bold"
  ) +
  annotate(
    "text",
    x = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 1),
    y = 19,
    label = "Hacked\nAction A",
    color = "#0072B2",
    size = 3.5,
    hjust = 1.3,
    fontface = "bold"
  ) +
  annotate(
    "text",
    x = management_scenarios_hacked |>
      pluck("habitat_quality_mean", 2),
    y = 19,
    label = "Hacked\nAction B",
    color = "#2C5F41",
    size = 3.5,
    hjust = -0.3,
    fontface = "bold"
  )

# Table of Summary Statistics
metric_labs <- c(
  "R2" = "R^2",
  "R2 adjusted" = "{R^2}_{adjusted}",
  "AIC wt" = "{AIC}_{wt}",
  "AICc wt" = "{AICc}_{wt}",
```

```
  "BIC wt" = "{BIC}_{wt}",
  "Sigma" = "\\sigma"
) |>
  map_chr(~ glue::glue("${.}$"))

performance_table <-
  performance::compare_performance(
    model_initial,
    model_overfitted,
    rank = TRUE
  ) |>
  select(-Model) |>
  mutate(
    Name = stringr::str_replace(
      Name,
      "model_overfitted",
      "Overfitted Model"
    ) |>
      stringr::str_replace(., "model_initial", "Initial Model")
  ) |>
  mutate(across(-c(Name), ~ round(.x, digits = 2))) |>
  pivot_longer(-Name) |>
  pivot_wider(names_from = Name, values_from = value) |>
  mutate(name = str_replace(name, "_", " ")) |>
  mutate(name = recode(name, !!!metric_labs)) |>
  mutate(name = vec_fmt_markdown(name)) |>
  gt(rowname_col = "name") |>
  text_transform(gt::md, cells_row_groups()) |>
  fmt_markdown(columns = name, rows = contains("$")) |>
  tab_header(
    title = "Stage 2: Model Fishing",
    subtitle = glue::glue(
      "The modeller compares the two models ",
      "and chooses the overfitted model based on these statistics."
    )
  ) |>
  tab_style(
    style = cell_fill(color = "#D55E00"),
    locations = cells_body(columns = "Overfitted Model", rows = everything())
  ) |>
  tab_style(
```

```
    style = cell_fill(color = "#E69F00"),
    locations = cells_body(columns = "Initial Model", rows = everything())
  ) |>
  tab_style(
    style = cell_text(color = "#D55E00"),
    locations = cells_column_labels(columns = "Overfitted Model")
  ) |>
  tab_style(
    style = cell_text(color = "#E69F00"),
    locations = cells_column_labels(columns = "Initial Model")
  ) |>
  tab_style(
    style = cell_text(
      font = google_font("Chivo"),
      size = "medium",
      weight = "bolder"
    ),
    locations = cells_column_labels()
  ) |>
  tab_style(
    style = cell_text(
      font = google_font("Cairo"),
      color = "black",
      size = "medium",
      weight = 500
    ),
    locations = cells_body()
  ) |>
  tab_style(
    style = cell_text(
      color = "black",
      font = google_font("Cairo"),
      size = "medium",
      weight = 400
    ),
    locations = cells_stub()
  ) |>
  tab_style(
    style = cell_text(
      color = "black",
      font = google_font("Roboto Condensed"),
```

```
      size = "large",
      weight = 400
    ),
    locations = cells_title(groups = "title")
  ) |>
  tab_style(
    style = cell_text(
      color = "black",
      font = google_font("Cairo"),
      size = "medium",
      weight = 400
    ),
    locations = cells_title(groups = "subtitle")
  ) |>
  cols_width(stub() ~ px(170), everything() ~ px(100)) |>
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_body(
      rows = name == "Performance Score",
      columns = "Overfitted Model"
    )
  ) |>
  tab_stub_indent(rows = name != "Performance Score", indent = 5)

# ----- Construct Patchwork Plot -----

tmp <- tempfile(fileext = ".png")
gtExtras::gtsave_extra(
  performance_table,
  tmp,
  zoom = 2,
  expand = 0,
  vwidth = 420,
)
table_png <- png::readPNG(tmp, native = TRUE)

patch <- (p1 + table_png) + plot_layout(widths = c(2, 1))
combined_plot <- patch /
  p3 +
  plot_annotation(tag_levels = c("A")) +
  plot_layout(heights = c(2, 3))
```

```
ggsave(
  filename = here::here("synthetic_example_QRPs.pdf"),
  device = grDevices::cairo_pdf,
  combined_plot,
  width = 17,
  height = 12,
  dpi = 600
)
```

D Supporting Information for the "Adaptive Preregistration" Study

Supporting Information for Chapter 5. The project's Open Science Framework (OSF) page can be found at osf.io/sq5j4, within which the linked material described below is located.

D.1 Modelling Workflow Identification

D.1.1 Task 1: Identify Scientific Workflows for Ecological Modelling

The goal of this analysis task was to identify scientific workflows for ecological modelling, both in practice, and idealised, with the broader aim of identifying a workflow that describes the modelling cycle that should underpin the structure and content of a preregistration template for ecological modelling. We conducted a literature review on 'good modelling practice' and structured decision making in ecological management, coding distinct tasks in the modelling workflow into 'phases,' 'steps' and 'sub-steps' using the qualitative data analysis software ATLAS.ti Scientific Software Development GmbH (2019). See the documents below, stored on the OSF at osf.io/tz5da:

- `PRT-WorkshopCoding - Memo Manager.csv` contains descriptions of modelling workflow phases derived from Modelling Workflow Identification Task 1 analysis.
- `references.bib` contains all references informing Task 1 Modelling Workflow Identification.

We then synthesised a preliminary idealised scientific workflow describing ecological modelling (osf.io/5w4ms/files/fgd23, osf.io/v7kjh). This work informed Workshop Activity

1, whereby participants collated and categorised modelling activities from their personal modelling workflows under the main phases of this workflow. See Appendix D.1 and Section 5.2.3 of the manuscript for workshop details.

D.1.2 Task 2: Describe a General Workflow for Ecological Modelling

Participant responses in workshop discussions were analysed using a combination of inductive and deductive coding using the qualitative coding analysis software ATLAS.ti (2019). Each decision-point from the personal modelling workflows in Workshop Activity 1 were coded as belonging to one of these phases, steps and sub-steps. We revised the coding structure according to patterns and themes identified across multiple personal workflows. For example, we added the final phase "model analysis", which aimed to capture the fact that analysts usually present the results of their modelling to clients, decision-makers or other stakeholders that must interpret the evidence and are responsible for making management decisions. In this way, we identified a common or generalised workflow from different types of applied ecological modelling projects that synthesised both idealised norms and norms of practice. This final workflow underpinned our draft preregistration template used in the case study.

The documents relevant to this analysis are stored on the OSF at osf.io/tz5da:

- `PRT-WorkshopCoding.xml` contains the entire Atlas.TI database of coded workshop outputs in `.xml` format.
- `individual_sheets_FW.csv` contains decision-steps identified in individual workflow worksheets for fieldwork. Decision-steps are coded into phases and sub decision-points following Workshop Activity 1 group worksheets (see Appendix S2 for details).
- `individual_sheets_MDA.csv` contains decision-steps identified in individual workflow worksheets for modelling. Decision-steps are coded into phases and sub decision-points following workshop activity 1 group worksheets (see Appendix S2 for details).
- `PRT-WorkshopCoding - Quotation Manager.csv` Contains individual modelling workflow decision-steps coded into modelling phases and decision-points.

- `PRT-WorkshopCoding - Code Manager.csv` modelling phases and decision-steps and their descriptions, derived from Modelling Workflow Identification Task 1 and Task 2 analysis.

- **Figure S1 – Coded Workshop Outputs** (See attached file "Figure-S1-coded-workshop-outputs.pdf"). Numbered boxes correspond to phases, steps and sub-steps of the modelling process. Each different phase of the modelling process is assigned a different colour. Coloured boxes are components of the modelling process that are synthesised from a combination of literature and workshop Task 1, with accompanying text describing the phase, step or sub-step. Grey boxes correspond to verbatim text from Task 1 workflows (indicated by the  symbol) or from the literature (indicated by the symbol), representing different modelling activities. The relationship of each node in the diagram is represented by arrows. Dashed arrows represent modelling activities that we have categorised as belonging to the associated modelling phase, step, sub-step, or activity. Whereas solid lines represent cases where the relationship is drawn from existing taxonomies of the modelling process reported in the literature, with directionality indicating hierarchy. The ordinal numbering of the modelling component corresponds to the phase, step and sub-step in our taxonomy of the modelling process.

D.2 User Research Workshop Materials

- Introductory Presentation 1: osf.io/wezy3
- Introductory Presentation 2: osf.io/egjwa

D.2.1 Activity 1 – Identifying Modelling Workflows

Note that this activity was run in two separate groups, with one group focusing on field work, and one focusing on modelling. Elements of the field work activity responses were collated into the analysis of the modelling activity outputs.

Materials

Workshop attendees were provided a 'Project Workflow and Decision Sequence' worksheet (osf.io/5w4ms/files/eza24) for individually describing decision steps and alternatives taken in a recent or memorable study. Next, each group collated their decision-steps from their

individual worksheets onto group worksheets for modelling (osf.io/5w4ms/files/fgd23) and for field studies (osf.io/5w4ms/files/w68j) with accompanying handouts describing workflow phases and steps (modelling: osf.io/5w4ms/files/v7kjh; field work: osf.io/5w4ms/files/uayt3).

D.2.2 Activity 2 – Identifying Challenges and Solutions to Preregistration for Modelling in Ecology

Materials

Activity 2 consisted of facilitated group discussions based on prepared questions and talking points informed by recent debates about preregistration in the literature, for example:

- What should a research workflow look like for a modeller in an applied ecological setting who wishes to preregister their work?
- What barriers are there for preregistering model-based research?
- How do we accommodate the iterative cycle of model development into the preregistration process?
- At what stage in a modelling or structured decision-making process should preregistration begin?
- Should we, and how can we, change the medium of the template and/or archiving platform to accommodate any roadblocks?
- How could we change the medium of both the template and the archiving platform to accommodate these issues?
- What procedural and or technical solutions could we implement to address these roadblocks to preregistration?
- What decision-tools could be used to aid in completing a preregistration?
- What sort of parsimony should we aim for, in terms of the resolution of the templates and therefore their applicability across different methodologies?

D.3 Adaptive Preregistration User Guide

The online user guide is reproduced below, excluding the template, which is presented separately in Section D.4. The guide is archived on Zenodo at:

> Gould, E., Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H., Vivian, L., Good, M., Duncan, D. H., Rumpff, L., & Fidler, F. (2026). EcoConsPreReg: A Guide to Adaptive Preregistration for Model-Based Research in Ecology and Conservation (v1.0.2.9001). Zenodo. https://doi.org/10.5281/zenodo.10807029

The source code is located at github.com/egouldo/EcoConsPreReg.

ADAPTIVE PREREGISTRATION FOR MODEL-BASED RESEARCH

A guide and template for preregistering ecological models

Table of contents

About this guide 1

Part I: Researcher Degrees of Freedom 2

From *Researcher Degrees of Freedom* to *Questionable Research Practices* 3
- Researcher Degrees of Freedom . . . 3
- Questionable Research Practices . . . 3
- Preregistration mitigates against researcher degrees of freedom . . . 4
- References . . . 5

Part II: Adaptive Preregistration 7

A primer on *Adaptive Preregistration* 8
- Background & Motivation . . . 8
- Iterative Cycle of Model Development and Preregistration . . . 9
- Adaptive Preregistration . . . 10
  - Preregistering Flexibility . . . 10
  - Interim Preregistrations . . . 11
- Documenting adaptive preregistration transparently . . . 11
  - The *Checking Problem* . . . 11
- Implementing *Adaptive Preregistration* with git and GitHub . . . 13
- References . . . 15

Implementing *Adaptive Preregistration* with git and GitHub 16
- Who is this guide for? . . . 16
  - Software and Set-up Requirements . . . 16
- Adaptive Preregistration using GitHub . . . 17
- How it works... . . . 18
  - ...Shifting from planning to execution... . . . 19
  - ... And back again. . . . 20
  - Flexible Strategies, Preliminary Analyses or Investigations . . . 21
- A step-by-step guide to implementing Adaptive Preregistration with GitHub . 21
  - The initial preregistration phase . . . 21
  - Incrementally Updating the Preregistration Plan . . . 23

Table of contents

Moving from planning to doing . . . 24
'Living preregistration' - tracking changes to the Preregistration Draft . 27
References . . . 31

Part III: Preregistration Template for Ecological Modelling **32**

Adaptive Preregistration for Modelling **33**
Problem Formulation . . . 33
Model Context and Purpose . . . 33
Scenario Analysis Operationalisation . . . 37
Define Conceptual Model . . . 37
Choose elicitation and representation method . . . 38
Explain Critical Conceptual Design Decisions . . . 38
Model assumptions and uncertainties . . . 38
Identify predictor and response variables . . . 39
Define prior knowledge, data specification and evaluation . . . 39
Conceptual model evaluation . . . 42
Formalise and Specify Model . . . 43
Model class, modelling framework and approach . . . 43
Choose model features and family . . . 43
Describe *approach* for identifying model structure . . . 45
Describe parameter estimation technique and performance criteria . . . 45
Model assumptions and uncertainties . . . 47
Specify formal model(s) . . . 47
Model Calibration, Validation & Checking . . . 48
Model calibration and validation scheme . . . 48
Implementation verification . . . 49
Model checking . . . 50
Model Validation and Evaluation . . . 51
Model output corroboration . . . 52
Choose performance metrics and criteria . . . 53
Model analysis . . . 53
References . . . 56

About this guide

This book/site provides a guide to implementing *Adaptive Preregistration for Model-Based Research*, with a focus on ecological modelling as it is used in both basic and applied contexts. This site covers the following topics:

1. Background on researcher degrees of freedom, the risk of questionable research practices (QRPs), and introduction to preregistration as a tool for mitigating QRPs and improving transparency.
2. Adaptive Preregistration
    1. *Overview*: What is adaptive preregistration, why do we need it for registering modelling analyses, and what are its defining features?
    2. *A practical guide*: How to implement adaptive preregistration using git and GitHub.
3. Preregistration template for model-based research in ecology and conservation.

This practical guide accompanies the paper:

> Gould, E., Jones, C.S., Yen, J.D.L., Fraser, H.S., Wootton, H., Good, M.K., Duncan, D.H., Hauser, C.E., Wintle, B.C., Rumpff, L. (in prep., 2026) "But I can't preregister my research: Improving the reproducibility and transparency of ecological modelling with adaptive preregistration". EcoEvoArXiv [Preprint]. https://doi.org/10.32942/X2GW66

Please cite this guide as:

> Gould, E., Jones, C.S., Yen, J.D.L., Fraser, H.S., Wootton, H., Good, M.K., Duncan, D.H., Hauser, C.E., Wintle, B.C., Rumpff, L. & Fidler, F. (2026). EcoConsPreReg: A Guide to Adaptive Preregistration for Model-Based Research in Ecology and Conservation. Zenodo. https://doi.org/10.5281/zenodo.19064144

Part I: Researcher Degrees of Freedom

From *Researcher Degrees of Freedom* to *Questionable Research Practices*

Researcher Degrees of Freedom

Throughout the course of model development and analysis researchers make numerous decisions. These alternative decisions are termed 'researcher degrees of freedom' (Simmons et al., 2011; Wicherts et al., 2016). These decisions are usually subjective and arbitrary yet methodologically and substantively acceptable (Wicherts et al., 2016), however may collectively significantly influence the final 'analysis strategy' (*sensu* Hoffmann et al., 2021) or 'specification' (*sensu* Simonsohn et al., 2020), and consequently the research findings and conclusions drawn from that analysis (Desbureaux, 2021).

Alternative decisions about model assumptions and implementation can lead researchers down very different analytical paths that result in 'parallel' or 'adjacent' worlds with different interpretations, and potentially conflicting conclusions originating from a single underlying dataset (Liu et al., 2020; Young & Stewart, 2021). The set of all (plausible or reasonable) possible decisions that a researcher could make over the entire analysis from beginning to end has been termed collectively as the 'garden of forking paths' (Gelman & Loken, 2013) or the 'multiverse', which consists of multiple universes of plausible analysis strategies (Steegen et al., 2016).

Researcher degrees of freedom are not inherently problematic, but rather are a product of the uncertainty that besets model development and analysis, whether due to incomplete theoretical understanding of underlying ecological processes, or because modelling is an act of simplification and abstraction; there is always some degree of choice about which ecological processes and phenomena should be included in the model, how they should be represented (Babel et al., 2019; Getz et al., 2017). Further uncertainty and multiplicity of analysis strategies stems from uncertainty in data pre-processing, model, parameter, and method uncertainty (Hoffmann et al., 2021).

Questionable Research Practices

When researchers are faced with ambiguous decisions where there is no single decision or strategy acceptable according to scientific standards and consensus, we are likely to select the decision that results in the most 'favourable' finding with "convincing

self-justification" (Hoffmann et al., 2021; Simmons et al., 2011). Researchers routinely execute and explore alternative analysis pathways in service of identifying the most favourable (i.e. publishable) result, selectively reporting only one or several preferred results (Liu et al., 2020; Simmons et al., 2011).

Importantly, as Gelman and Loken note (Gelman & Loken, 2013) this behaviour is not a consciously nefarious 'fishing expedition', rather, because researchers are motivated actors in a 'publish or perish' culture who are susceptible to cognitive biases and are willing to change their modelling assumptions when the results conflict with desired outcomes – sincerely and self-convincingly believing that the more favourable model is superior (Young & Stewart, 2021).

Thus researcher degrees of freedom precipitate questionable research practices when they remain both undisclosed and exploited opportunistically (Bakker et al., 2018). Undisclosed and opportunistic use of researcher degrees of freedom artificially increases the probability of false positive results, inflates effect size estimates and ultimately reduces the replicability of published findings while encouraging overconfidence in the precision of results and exacerbating the risk of accepting and propagating false facts (Hoffmann et al., 2021; Nissen et al., 2016; Roettger, 2019; Wicherts et al., 2016). Emerging evidence suggests that rates of QRPs in ecology and evolutionary biology are similar to rates observed in other disciplines beset by reproducibility crises (Fraser et al., 2018).

Preregistration mitigates against researcher degrees of freedom

Preregistration is an open-science practice adopted in scientific disciplines that have begun to confront the 'reproducibility crisis', and aims to improve research transparency and mitigate questionable research practices within a study by distinguishing between what is a genuine a priori planned analysis, and what is not (Parker et al., 2019). Preregistration requires the methods and analysis plan of a future study are registered in a secure and publicly accessible platform prior to any data collection, and after submission it can no longer be altered.

Despite enthusiastic uptake of preregistration in some scientific disciplines, there has been little uptake of pre-registration in ecology and conservation, and many modellers and researchers engaged in non-hypothesis testing research both within and outside of ecology have eschewed preregistration on the grounds that preregistration and existing templates are irrelevant because they are focused on null-hypothesis significance testing (e.g. McDermott et al., 2021; Prosperi et al., 2019).

Instead of eschewing preregistration completely, this guide is premised on the idea that preregistration can and should be used in model-based research in ecology and conservation, but its particular form should reflect the norms and practice of the research context in which it is applied, in order to adequately restrict researcher degrees of freedom that may uniquely emerge within a domain-specific methodology. The work presented

in this guide aims to translate preregistration into model-based research within ecology and conservation by proposing an expanded view of preregistration we term *Adaptive Preregistration*. We also present a template for preregistering model-based research in ecology and conservation.

References

Babel, L., Vinck, D., & Karssenberg, D. (2019). Decision-making in model construction: Unveiling habits. *Environmental Modelling & Software*, *120*, 104490. https://doi.org/10.1016/j.envsoft.2019.07.015

Bakker, M., Veldkamp, C. L. S., Assen, M. A. L. M. van, Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. T., & Wicherts, J. M. (2018). Ensuring the quality and specificity of preregistrations. *PsyArXiv*. https://doi.org/10.31234/osf.io/cdgyh

Desbureaux, S. (2021). Subjective modeling choices and the robustness of impact evaluations in conservation science. *Conservation Biology*, *35*(5), 1615–1626. https://doi.org/10.1111/cobi.13728

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Getz, W. M., Marshall, C. R., Carlson, C. J., Giuggioli, L., Ryan, S. J., Romañach, S. S., Boettiger, C., Chamberlain, S. D., Larsen, L., D'Odorico, P., & O'Sullivan, D. (2017). Making ecological models adequate. *Ecology Letters*, *21*(2), 153–166. https://doi.org/10.1111/ele.12893

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A. L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Liu, Y., Althoff, T., & Heer, J. (2020). Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376533

McDermott, M. B. A., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., & Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, *13*(586). https://doi.org/10.1126/scitranslmed.abb1655

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife*, *5*, e21451. https://doi.org/10.7554/eLife.21451.001

Parker, T., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, *33*(4), 747–750. https://doi.org/10.1111/cobi.13342

Prosperi, M., Bian, J., Buchan, I. E., Koopman, J. S., Sperrin, M., & Wang, M. (2019). Raiders of the lost HARK: A reproducible inference framework for big data science. *Palgrave Communications*, *5*(1). https://doi.org/10.1057/s41599-019-0340-8

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1). https://doi.org/10.5334/labphon.147

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Aert, R. C. van, & Assen, M. A. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Young, C., & Stewart, S. A. (2021). Multiverse analysis: Advancements in functional form robustness. *Cornell University, Ithaca, NY. Unpublished Working Paper.*

Part II: Adaptive Preregistration

A primer on *Adaptive Preregistration*

Background & Motivation

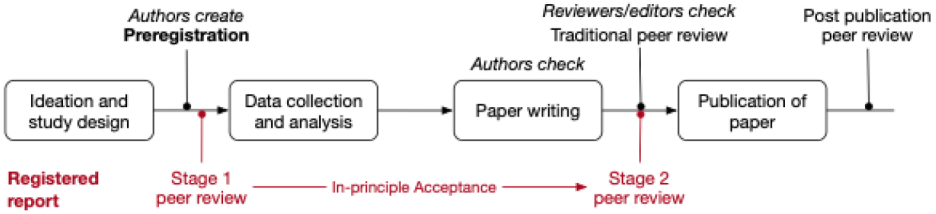


Figure 1: Preregistration as a component of the registered report publication pathway, as illustrated by (Pu et al., 2019).

The iterative nature of model development is a significant barrier to adopting preregistration for model-based research. Translating the preregistration process to a modelling research context is challenging because the iterative nature of model development conflicts with the inner logic of existing preregistration templates that presume a linear research workflow. The process of preregistration as it is currently implemented is tailored towards the hypothetico-deductive model of the scientific method. As such, the process of preregistration is completely distinct from, and precedes the implementation of the analysis plan laid out in the preregistration – analytic decisions are completely independent from the data.

In fact, preregistration has been defined as "the action of confirming an unalterable version of one's research plan prior to collecting data" (Pu et al., 2019) or prior to analysing the data (Mertens & Krypotos, 2019). In this process, researcher sequentially shifts from ideation, to study/analysis design and preregistration, to collecting and analysing the data, writing the report or manuscript, to documenting the research and publishing (Figure 1). The iterative nature of model development, however, inherently precludes this model of preregistration from being applied to model-based research in ecology and conservation.

A primer on Adaptive Preregistration

Iterative Cycle of Model Development and Preregistration

Current preregistration practices based on NHST-focussed accounts of questionable research practices prescribe the principle of 'decision-independence', where analysis decisions are made independently of and *a priori* to analysing the data (Gelman & Loken, 2013; Srivastava, 2018). However, the modelling process may genuinely violate the principle of 'decision-independence' because it is non-linear, iterative and usually generates many interim versions of the model preceding publication. During modelling a researcher will almost always engage in state-dependent decision-making — that is, decisions about whether to alter the model in a subsequent round of model development depend on the results of model testing, evaluation and analysis at the end of the previous iteration of model development. These practices aren't necessarily questionable in and of themselves, but do have the *potential* to become questionable if they are exploited in service of artificially inflating the accuracy or precision of the model, its predictions or evaluation tests to the effect that the model is perceived to be more credible than it would be if that practice did not occur – essentially data-dependent decisions that lead to consumers of the model placing unsubstantiated belief in the reliability, validity and utility of the model and its outputs are questionable and can be mitigated through the practice of preregistration.

Very often in modelling, there are decisions that are necessarily dependent on the outcome of previous analytic decisions in the modelling workflow (Liu et al., 2020). Firstly, preliminary or investigatory analyses might need to be conducted before being able to specify future decision-steps in the analysis plan. For example, modellers might need to check the distribution of particular variables in order to determine how to specify the model, perform assumption checks, or check for collinearity or spatial autocorrelation. Secondly, data-dependent decisions from model checking may justifiably result in changes to either the analysis or to interpretations of the mode (Srivastava, 2018). For instance, inspection of the residuals and other model checking tests may force the modeller to return to earlier decision-points and change the planned analysis or model. Other times, the planned analysis and specified model may simply not converge, or the model is saturated and runs out of variation to apportion such that the model and or the model fitting algorithm must be respecified and re-implemented. Other times, the modelling process itself may generate knowledge and learning about the system, and the model structure changes throughout the process of model development.

Finally, sometimes there are some decisions that are too difficult to anticipate, or simply cannot be made in advance (Srivastava, 2018). This is particularly true for decisions occurring at the later phases of the modelling construction and development process, where downstream decisions depend on the outcomes of earlier decisions and outcomes of modelling analyses. Some decisions might not be able to be specified on the first attempt at writing the preregistration until the model is fully or at least close to fully specified, for example, specifying precisely what and how sensitivity analyses or uncertainty analyses will be conducted.

As illustrated above, the adaptive nature of analysis during the model development process contravenes the central principle of preregistration – decision-independence and is fundamentally incompatible with the sequential, linear process of traditional preregistration where the analysis plan is fully specified prior to observing and analysing the data.

Adaptive Preregistration

In order to address the disjuncture between the iterative nature of model development, and the single-use, deterministic preregistration template format, we take an 'expanded view of preregistration', that fits with the concept of 'adaptive preregistration' proposed by Srivastava (2018). The modeller proceeds through the model development process, shifting from ideation, preregistration, execution of the analysis and back to ideation and preregistration again. As they progress, the modeller implements parts of the analysis plan that were specified in the previous preregistration step, with the outcomes of that analysis step either informing the specification of the next analysis (Figure 4) or else resolving the preregistered decision.

This view of preregistration breaks with the current model of preregistration, wherein the author writes a single deterministic preregistration containing a rule for every decision, and sequentially shifts from ideation, preregistration to execution of the plan. Two features define adaptive preregistration:

1. It contains 'plans to deploy flexible strategies' (Srivastava, 2018) wherein the author can supply heuristic consisting of multiple different analysis or modelling strategies whose execution depends on the outcome of previous decision-points or analyses, for example, decision-trees could be preregistered prior to analysing the data (Baldwin et al., 2022).
2. The preregistration may be iterative and consist of interim preregistrations that mark phases of modelling and analysis as different parts of the data are observed (Srivastava, 2018).

Preregistering Flexibility

When preliminary analyses, model checking or other data-dependent decisions need to be made, the modeller can specify a 'flexible strategy' (Figure 2) in their preregistration by:

- stating what quantity or outcome needs to be known to move forward with the modelling and analysis,
- explaining what test or analysis will be performed to obtain this quantity or outcome, and what parts of the data will be used in this analysis, and

- describing how the results will be interpreted, listing each potential decision and its trigger, where possible.

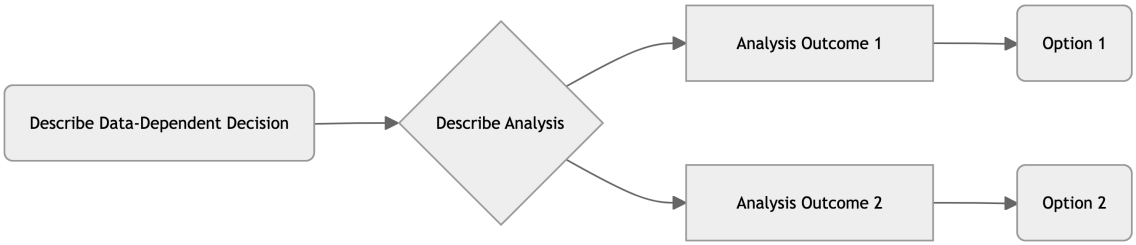


Figure 2: A general heuristic for registering flexible analyses.

Interim Preregistrations

The modeller follows an iterative process of preregistration. As they proceed through the model development process, they can shift from ideation, preregistration and execution of the analysis plan, to preregistration again. Depending on observed outcomes of the pre-specified decision-trees or registered flexible heuristics, interim preregistrations are created at multiple points in the model development process, each time there is an addition or amendment to the analysis plan.

Documenting adaptive preregistration transparently

The *Checking Problem*

Following publication of the final study, authors, reviewers, editors and readers may all engage in the process of checking a manuscript or published paper against its preregistration to verify that the study and analyses were conducted as specified in the preregistration. The process of preregistration is already cumbersome, because the dominant format of the preregistration is a single, static text-based document designed with the study authors in mind to very quickly insert the required analysis plan information, rather than to facilitate a comparison of the reported analysis and results against the planned analysis in the preregistration. In their current form, they are almost exclusively exist as "write-only media" (Pu et al., 2019). The task of model checking is further complicated by the non-linear and iterative nature of model development, and by our proposed adaptive preregistration methodology, because in most modelling studies there will likely be interim preregistrations, such that there are multiple versions of the preregistration to check against the completed analysis.

A primer on Adaptive Preregistration

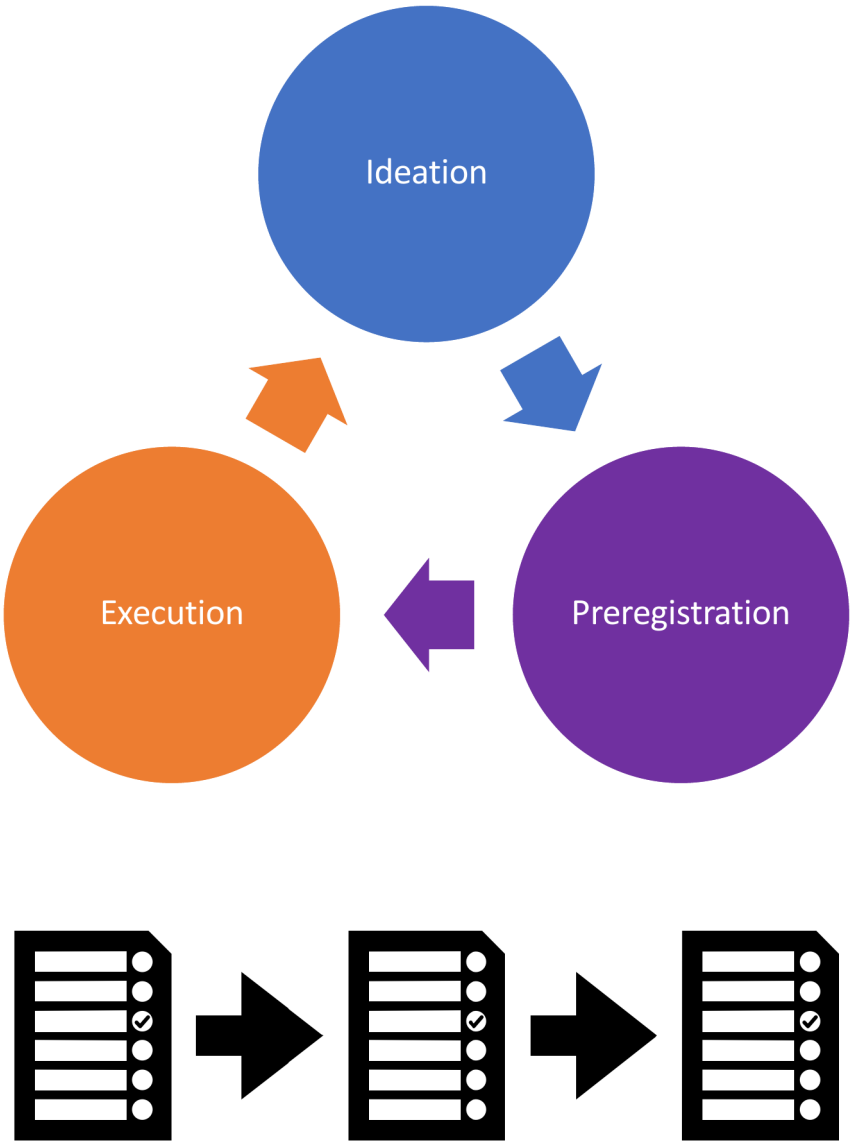


Figure 3: Adaptive preregistration can cycle through the specification, execution and ideation phases multiple times in adaptive preregistration, with interim versions of the preregistration being generated at each preregistration point.

Implementing *Adaptive Preregistration* with git and GitHub

Given the iterative nature of adaptive preregistration then, the current convention of the single-use, deterministic and static text-based document format for preregistration is inadequate. Interim versions of a preregistration must be time-stamped and ideally version-controlled, such that it is clear not just that two or more versions differ from each other, and in what order they were created, but so that it is explicitly clear in how and why they differ. To facilitate preregistration checking of flexible modelling and analysis strategies, we must link the results of both final and preliminary analyses back to the specified strategy: the link between a particular triggered analysis decision, and the analysis outcome that triggered the decision must be explicit. Additionally, if the results of some preliminary analyses cause revisions to the model itself or a different modelling procedure or approach to be selected than was originally or previously planned, then the trigger for this decision must be explicitly linked to the next version of the preregistration.

We propose leveraging the features of git and GitHub to facilitate implementation both adaptive preregistration in a transparently documented manner (Figure 4), and provide a step-by-step guide.

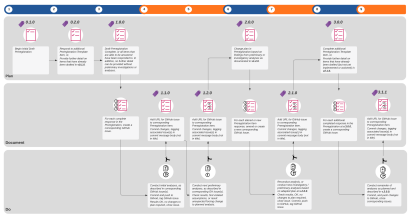


Figure 4: Adaptive Preregistration for model-based research using GitHub.

References

Baldwin, J. R., Pingault, J. B., Schoeler, T., Sallis, H. M., & Munafò, M. R. (2022). Protecting against researcher bias in secondary data analysis: Challenges and potential solutions. *European Journal of Epidemiology*, *37*(1), 1–10. https://doi.org/10.1007/s10654-021-00839-0

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Liu, Y., Althoff, T., & Heer, J. (2020). Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376533

Mertens, G., & Krypotos, A. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*(1), 338–352. https://doi.org/10.5334/pb.493

Pu, X., Zhu, L., Kay, M., & Conrad, F. (2019). Designing for preregistration in practice: Multiple norms and purposes. *Open Science Framework*. https://doi.org/10.31219/osf.io/w5ked

Srivastava, S. (2018). Sound inference in complicated research: A multi-strategy approach. *PsyArXiv*. https://doi.org/10.31234/osf.io/bwr48

Implementing *Adaptive Preregistration* with git and GitHub

Who is this guide for?

This guide assumes a basic understanding and knowledge of version-control tools, like git and/or GitHub. However, we have provided some background knowledge, links and references throughout this guide so that any modellers or researchers can set up and implement adaptive preregistration with git and GitHub for a new project. As a starting point, we direct readers to Braga et al. (2023), which provides an excellent introductory overview of the GitHub's functionality.

For those with basic familiarity of git and GitHub, we use a shared repository model for collaboration. If you are not yet familiar with GitHub features, such as pull requests and branches, we have linked to the relevant GitHub documentation pages where further background reading might be necessary.

We suggest some tooling that assumes use of the R statistical programming language (R Core Team (2024), see Section), since that is the most common programming language in the fields of ecology (Gao et al., 2025), however the general approach and principles described in this document is language agnostic, and does not require the use of R.

Software and Set-up Requirements

Follow the instructions at https://docs.github.com/en/get-started/git-basics/set-up-git to setup the following:

- git installation on your computer (Windows, Mac, linux supported)
- GitHub user account
- Configure authentication with GitHub from git

After you have installed and set-up the required software and GitHub account, you will need a git/GitHub repository (see Quickstart for repositories) where your preregistration document and analysis files will be version-controlled.

Implementing Adaptive Preregistration *with git and GitHub*

> Adaptive Preregistration Primer
>
> Make sure you have read the primer on Adaptive Preregistration and background before preceding with this guide.

Adaptive Preregistration using GitHub

We propose leveraging both the version-control and the project management features of GitHub (www.github.com) as a tool for implementing adaptive preregistration. GitHub is a web-interface that utilises the version-control system, 'git', to store files, track changes, and enable collaboration on computer code (Braga et al., 2023). When a document is version-controlled using Git and GitHub, each version of a document can be time-stamped, and is assigned its own unique identifier, or commit-hash[1] so that chronological records of changes to a collection of files, stored in a 'repository', are made transparent (Braga et al., 2023; Bryan & Crossman, 2008).

Why GitHub for Adaptive Preregistration?

- Because GitHub documents are 'time-stamped' the genesis of the preregistration from one version to the next is apparent.
- Moreover, using the diff view of a document on GitHub, exactly how and where a document has changed between versions is clear - additions are coloured green, deletions are coloured red, and changes within a line are highlighted.
- The project management and communication features of GitHub, such as GitHub Issues,
- tagging and release

GitHub users can collaborate by contributing, modifying and discussing existing code, reporting bugs and other problems, discover new code and data, and publish new code.

GitHub also integrates several communication features, such as GitHub Issues, GitHub Discussions, GitHub Pages, which allows users to engage in discussions, to plan and collaborate on code, and publish information to a webpage (see Box 1 in Braga et al., 2023).

The preregistration document should be saved within the project repository that contains the data and modelling and analysis code, such that any results from analyses that cause the analysis plan to change can be linked and referenced in the updated version of the plan – thus facilitating the process of model checking. By providing a method for contrasting the actual analysis undertaken against the preregistered analysis, this provides a mechanism for allowing reviewers to properly evaluate the preregistration

[1] https://docs.github.com/en/free-pro-team@latest/github/getting-started-with-github/github-glossary#commit-id

Implementing Adaptive Preregistration *with git and GitHub*



Figure 1: 'Diff View' of changes made to a preregistration in process.

against the reported analysis. It thus provides a way of marking preregistered parts of a report from non-preregistered parts of the analysis.

How it works...

GitHub is used to track amendments to the preregistration – each time an additional template item is answered, or an alteration is made to the preregistration, the change should be *committed* to the repository. *Semantic versioning* is used to document each version of the preregistration. Semantic versioning[2] is an open-source software engineering practice used to document and and communicate the development of software, in a way that allows others to depend on it (Kitzes et al., 2018). We have created our own semantic versioning scheme to be used to specifically document the incremental version of the preregistration, in tandem with GitHub's tag and release feature, each interim preregistration is tagged and released with its semantic version number. All template items that can be completed should be completed before any data analysis or modelling proceeds. The preregistration should be completed until a) no further detail can be

[2]The semantic version is a 3-part numeric identifier separated by points, e.g.: `0.0.1`, wherein the first number sequences a major version of the preregistration, the second number sequences a minor change, and the third number sequences small changes or patches, yielding: `MAJOR.MINOR.PATCH`. See https://semver.org/ for further information. Further detail on how to use semantic versioning with adaptive preregistration is summarised in Table 1.

supplied or b) no additional template item can be answered until preliminary investigations or analyses are undertaken (Figure 2, steps 1 - 3).

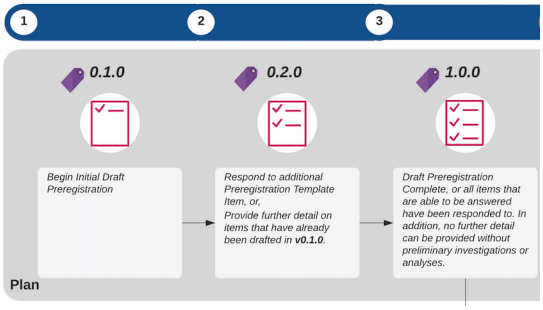


Figure 2: Initial Preregistration: All template items that can be answered should be completed prior to data analysis or modelling.

...Shifting from planning to execution...



Figure 3: Document: After the initial preregistration, document the initial preregistration using GitHub issues.

Before shifting from planning to the execution of the interim preregistration, each complete response in the interim preregistration should have a corresponding GitHub issue[3] created (Figure 3, step 3, 'Document'). Relevant discussion between analysts about how to proceed with the intended analysis or interpreting preliminary results are tracked within an issue's thread. Each issue is automatically assigned a unique number by GitHub. Any code or analysis outputs, such as figures, tables or other files, should be committed and tagged with the issue number, so that all analysis addressing that task is tracked within the issue's thread (Figure 4, step 4, 'Do'). The URL for the GitHub issue should then be added to the relevant preregistration item, the preregistration document should then be committed with the changes, and the version number of the preregistration updated in a new release.

[3] https://guides.github.com/features/issues/

Implementing Adaptive Preregistration *with git and GitHub*

... And back again.

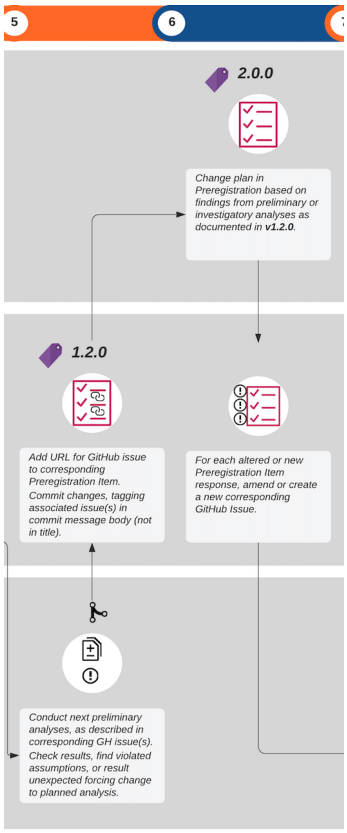


Figure 4: Adaptive Preregistration - When Plans Change

As the modelling proceeds, further detail on later phases of the analysis can be iteratively updated and preregistered. Where the results of model evaluation and analysis reveal that there are problems with the model, plans can be changed, and again the next phase can be preregistered. For example, if a researcher finds that assumptions are violated or other unexpected results occur that force a change to the planned analysis, the preregistration can be updated based on the findings of those analyses. When this occurs, the major version number of the preregistration document should be incremented (Figure 4, steps 5 - 6). Since the findings of the analyses are tracked within the relevant issue for that template item, and the issue is recorded in the preregistration itself, the trigger for the change in the plan is made explicitly clear.

For each altered or new preregistration item, the issue thread should either be updated, or a new issue created respectively (Figure 4, step 6). Those analyses are then either reconnected based on the revised plan or the new investigatory analyses are then conducted (Figure 4, step 7).

This process continues until the preregistration is complete, and the researcher can

continue to execute the plan as it has been fully described in the final version of the preregistration document (Figure 4, step 8 - 9).

Flexible Strategies, Preliminary Analyses or Investigations

When a researcher needs to conduct preliminary analyses or check parts of the data before committing to a particular decision about the model or analysis, they can specify a flexible strategy in their response to that preregistration item by:

- stating what needs to be known to move forward with the modelling and analysis and why the analysis is necessary
- explaining how the researcher will test this, and what parts of the data will be used in that analysis
- describing how the results will be interpreted, listing each potential decision and the analysis result that will trigger that decision, where possible.

A step-by-step guide to implementing Adaptive Preregistration with GitHub

The initial preregistration phase

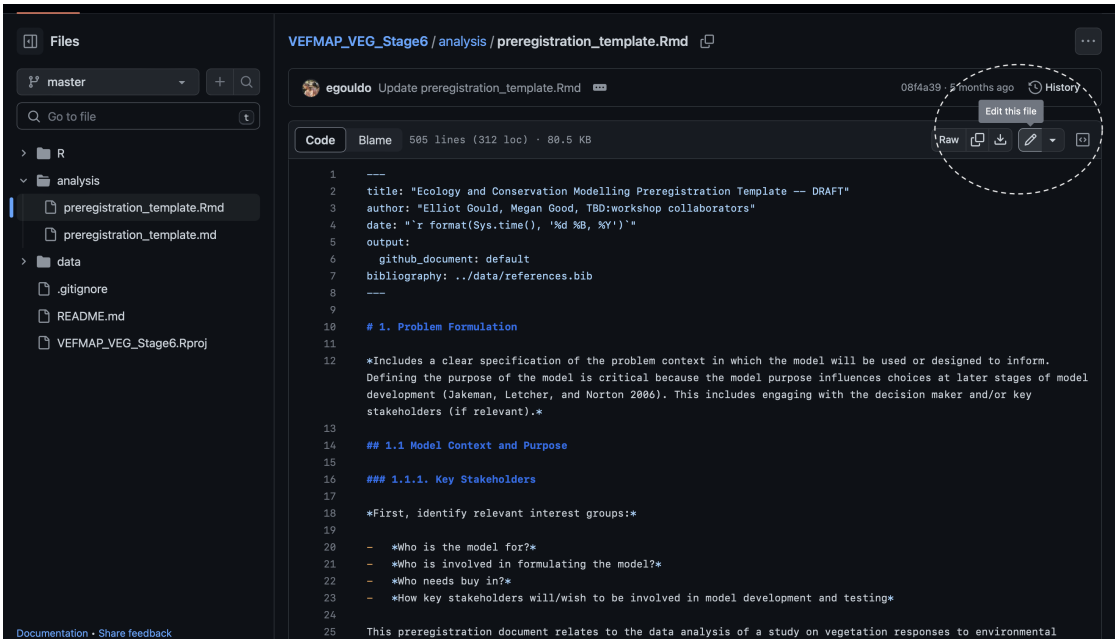
> Preregistration Template
>
> We have developed a template for preregistering ecological modelling studies. The template is available for use as a quarto template extension, which allows rendering to `.html`, `.pdf` or `.docx` file formats. Should you wish *not* to use quarto markdown as the format for your preregistration document, feel free to copy the relevant content into the text file format of your choice.
> Ensure you have committed a preregistration template to your project repository before proceeding with the following steps.

The first iteration of preregistration

1. Create a GitHub issue describing any tasks relevant to creation of the initial preregistration
2. Edit the document with your responses to the template items

- View our preregistration template document in GitHub here.
- In your own repository containing your preregistration document, click on the 'edit document' button:

Implementing Adaptive Preregistration *with git and GitHub*

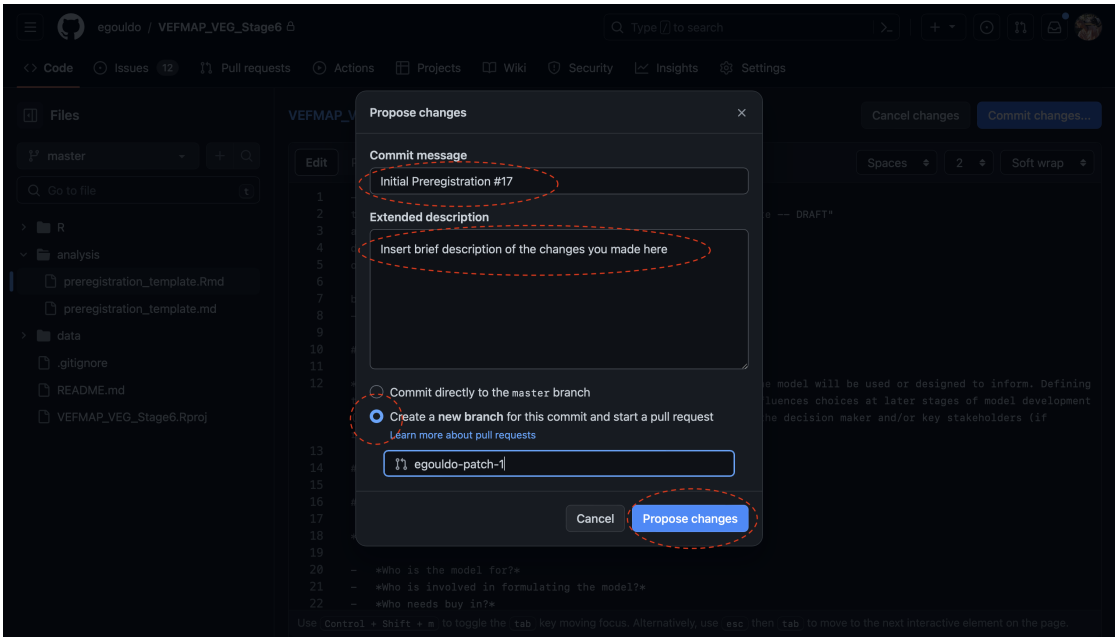


- You will be able to edit the document directly in GitHub, Edit away!

2. Commit changes

- If you are finished, or ready to save your changes by ‘committing’ them. There is a box called “commit changes” at the top right of the window, a pop-up window will appear, enter in the commit message box the following:
  - `Initial Preregistration #IssueNumber`. Tag the Preregistration GitHub issue number in the commit message (For this repository, it’s `#17`, however if you have only created the issue for the initial preregistration, then your issue number will be `#1`).
- Fill out the extended description box briefly describing the changes you just made
- Ensure that the radio button for creating a new branch is selected, an auto-generated branch name will appear, e.g. `egouldo-patch-1`.
- Click “Propose Changes” to start a pull request:

Implementing Adaptive Preregistration *with git and GitHub*



! Pull Requests & the GitHub flow model of collaboration

The approach to using GitHub in this document follows the GitHub flow model for collaboration on projects. However, if you are working independently rather than collaborating on your project in a team, you can instead skip the use of pull requests and branches, and select the ‘commit directly to the `default` branch’ radio button above.
Note that if using the GitHub flow model demonstrated in this guide, before proceeding to the next step, you will need to make sure that the pull request is reviewed, updated if necessary, and the changes merged into the default branch once the pull request has been approved.
See the ‘reviewing changes in pull requests’ guide for further details.

Incrementally Updating the Preregistration Plan

Moving from general to specific: Incrementally updating the preregistration

You will likely find that some items on the template cannot be answered just yet. One reason being that the level of detail required to adequately answer the template item might not be clear yet, and more thought and time is needed, or else you need to be further along the modelling process in order to adequately answer the preregistration item. Another reason is that downstream decisions might depend on some preliminary investigations of the data – for example, you might need to explore the shape of a measured variable before you can properly specify the full model.

In the latter situation, where possible, you can preregister your preliminary analysis: describing what the preliminary analysis entails, providing a short and relevant list of plausible outcomes, and describing for each of those possible outcomes how the dependent or down-stream decision-point will change depending on the outcome.

Using GitHub to update the plan

We will use GitHub to track amendments to the preregistration for the two situations described above (a. adding more detail, answering incomplete questions; b. Responding to findings of a preliminary analysis already described in your preregistration).

To make the changes to the preregistration document, follow the same instructions as for creating the initial preregistration above - simply edit the preregistration in GitHub, and commit the changes.

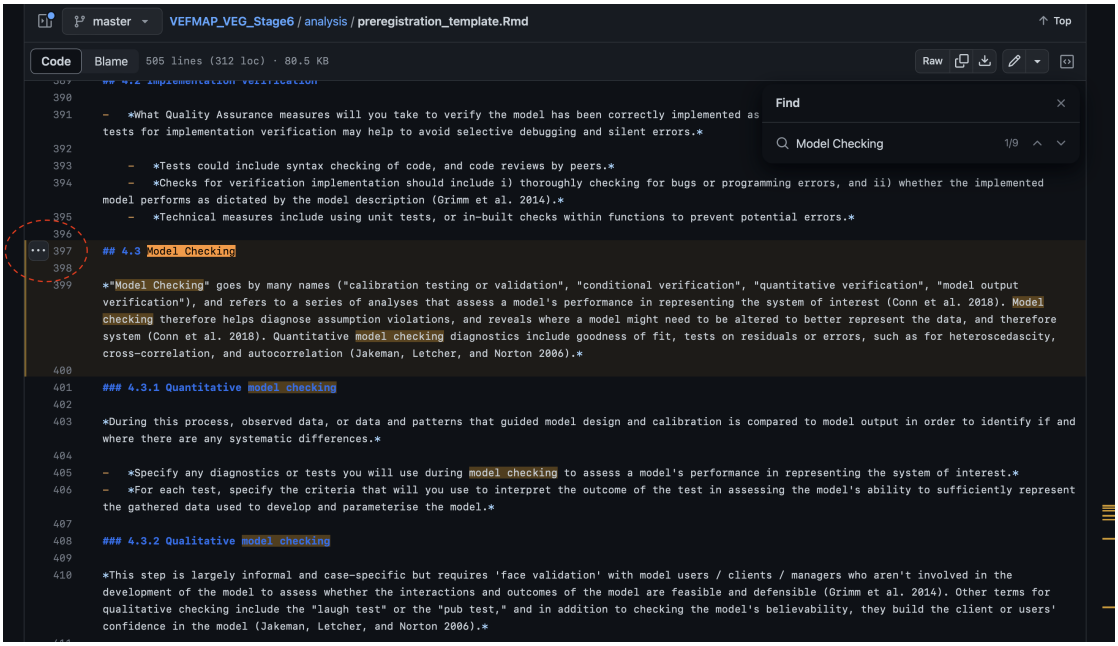
As for the situation where you are responding to findings of a preliminary analysis that you have already described in your preregistration, see the below sections on `Moving from planning to doing` and `Shifting back to planning from doing`.

Moving from planning to doing

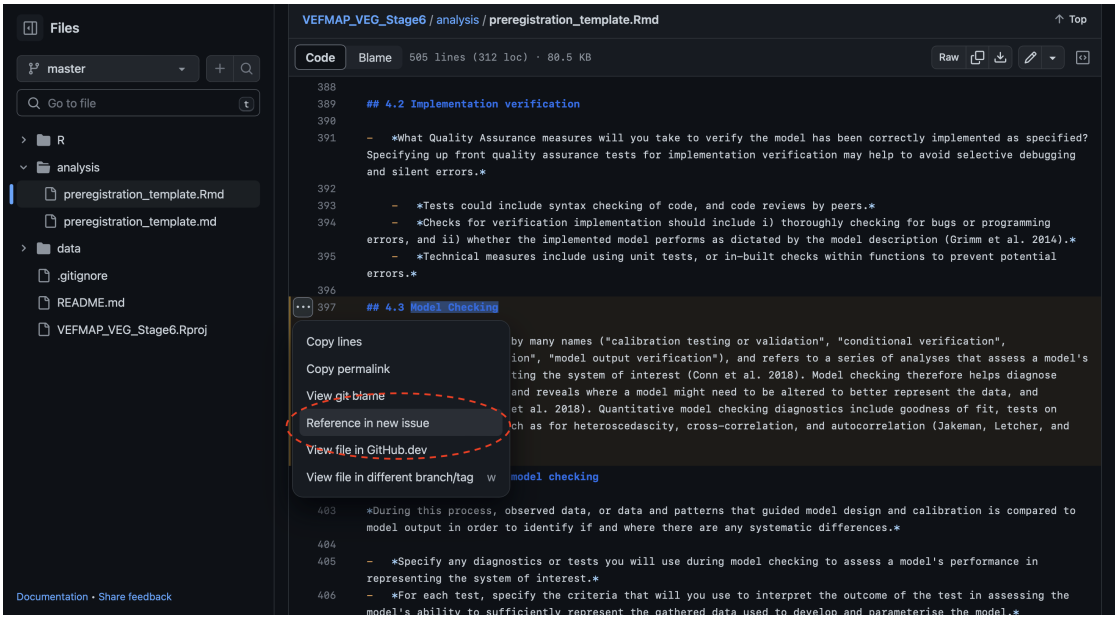
Before you begin to do or start anything you've planned within your preregistration, you **must** create a GitHub issue for that item, phase, or step in the modelling process, follow these steps:

1. View the preregistration document in GitHub, and click on the line-numbers on the left to highlight the preregistration template item text, as well as your response to the item. For example, let's say you want to start working the model checking, you would highlight the following in the document. Note that once you've highlighted some lines, a little box with an ellipsis appears on top of the lines numbers.

Implementing Adaptive Preregistration *with git and GitHub*

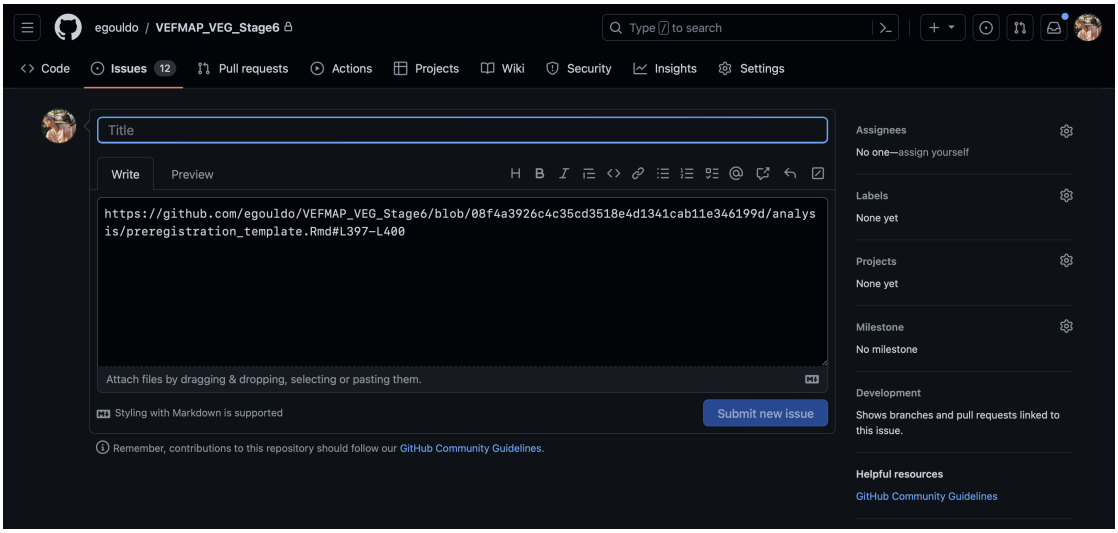


2. Click on the box with the ellipsis, and a pop-up dialogue box will appear with 4 options: `Copy lines`, `Copy permalink`, `View git blame`, `References in new issue`. Select the last option `Reference in new issue`.

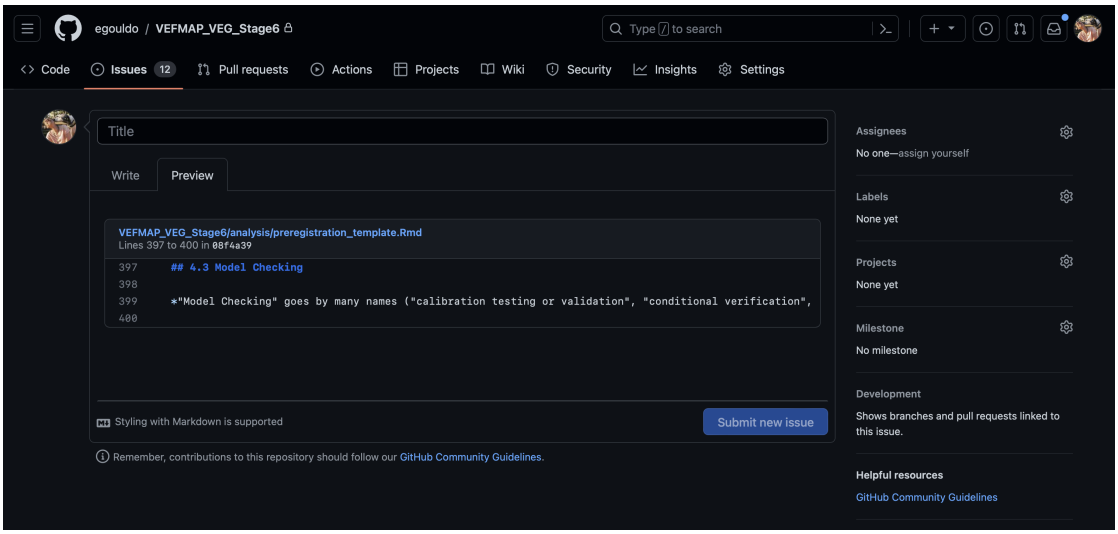


3. This will open up a new window with a new GitHub Issue for you to complete. Note that part of the issue description has already been filled out for you, this is a reference to the lines you just highlighted:

Implementing Adaptive Preregistration *with git and GitHub*



Which is cool, because if you open the `Preview` tab to view the preview of your issue, it will look like this:



Add any further information or technical details about going forward and working on this task, as necessary. Submit the issue to save. Any work, and subsequent commits on this topic should be tagged with the issue number. We now have a mechanism for linking your actual analysis and actions to the planned analysis in the preregistration document.

Implementing Adaptive Preregistration *with git and GitHub*

Dealing with unexpected outcomes, and changing plans: Shifting back to planning from doing

Sometimes assumptions are violated, or there are other unexpected outcomes of preliminary analyses, and you will need to change plans. How do we deal with this?

- Add the URL to the relevant GitHub issue under the corresponding Preregistration Item.
- Add any the URL for any linked commits to the Preregistration Item.
- Revise the Preregistration, where necessary by editing the Preregistration document in GitHub, as per previous instructions.
- Commit the changes with the following commit message header (or similar): `Preregistration Deviation #issue`.
- Describe the trigger for the Preregistration revisions in brief detail within the commit message body, adding the link to the relevant GitHub issue containing the tagged commits that triggered the revision.

'Living preregistration' - tracking changes to the Preregistration Draft

We will leverage GitHub's tagging and release feature in conjunction with Semantic Versioning to document and track changes to the preregistration document.

GitHub releases mark software iterations that you can release and deploy[4]. Within the context of adaptive preregistration, we can use releases to mark interim versions of the preregistration, as well as to take a snapshot of the state of the entire repository at that moment in time.

A summary of the types of preregistration changes, and the corresponding required version and GitHub features are presented in Table 1. Note that anything up until the complete preregistration should be additionally tagged as a 'pre-release' to communicate that the initial preregistration is not yet complete.

Using the `fledge::` R package for semantic versioning and to document preregistration changes

The package `fledge::` can be used right from the R console to bump the version number, make releases, and auto catalogue a chronological history of changes recorded a `NEWS.md` file that is updated and published with every release. A detailed guide to using `fledge::` can be found here https://fledge.cynkra.com/articles/fledge.html.

[4]Instructions for creating releases are here: https://help.github.com/en/github/administering-a-repository/managing-releases-in-a-repository

Implementing Adaptive Preregistration *with git and GitHub*

This package is not a mandatory component of the process, however can be a useful tool for implementing semantic versioning, tagging and releasing. All of these actions can be implemented in the GitHub browser with a point and click interface.

Table 1: Types of changes to the preregistration and corresponding actions and versioning adjustments

Preregistration Change or Action	GitHub Mechanism	Semantic Versioning
Changes preceding initial completion of preregistration	Tag and Mark as 'pre-release'	`Minor` increment. E.g. `0.2.0` The `Major` value should not exceed `0`. `fledge::bump_version("minor")`
Incremental Change, e.g. respond to new question.	Tag and Release	`Minor` increment. E.g. `0.2.0` The `Major` value should not exceed `0`. `fledge::bump_version("minor")`
Initial Preregistration complete. Note that 'complete' needs to be decided. Is complete after all preliminary and investigatory analyses have been done and the final Preregistration is complete, or is it the first iteration, where no more items can be filled out until some initial analyses are undertaken?	Tag and Release	`Major` increment to `1.0.0`. `fledge::bump_version("major")`; `fledge::finalize_version()`.
Major change to a complete Preregistration – no looking or analysing at data.	Tag and Release	`Major` increment `>1.0.0`. `fledge::bump_version("major")`; `fledge::finalize_version()`.

Preregistration Change or Action	GitHub Mechanism	Semantic Versioning
Linking implemented work to the Preregistration Item	Update Preregistration with GH Issue URL and all relevant commit hashes pertaining to that Preregistration Item + Commit. Tag and Release that preregistration commit.	`Minor` increment. `fledge::bump_version("minor")`.
Data-dependent changes with registered flexibility Adapting the Preregistration on the basis of results from preliminary or investigatory analyses	Update Preregistration with GH Issue Link and all relevant commit hashes pertaining to that Preregistration Item + Commit. Respond to dependent incomplete Preregistration Items or Update existing / already completed Preregistration Items + Commit. Tag and Release.	`Major` increment. `fledge::bump_version("major")`.
Minor mistakes or changes	Changes including small typos, or minor elaborations on existing specified plans	`Patch` increment. E.g. `0.1.2`. `fledge::bump_version("patch")`

References

Braga, P. H. P., Hébert, K., Hudgins, E. J., Scott, E. R., Edwards, B. P. M., Sánchez Reyes, L. L., Grainger, M. J., Foroughirad, V., Hillemann, F., Binley, A. D., Brookson, C. B., Gaynor, K. M., Shafiei Sabet, S., Güncan, A., Weierbach, H., Gomes, D. G. E., & Crystal-Ornelas, R. (2023). Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution. *Methods in Ecology and Evolution*, *14*(6), 1364–1380. https://doi.org/10.1111/2041-210X.14108

Bryan, B. A., & Crossman, N. D. (2008). Systematic regional planning for multiple objective natural resource management. *Journal of Environmental Management*, *88*(4), 1175–1189. https://doi.org/10.1016/j.jenvman.2007.06.003

Gao, M., Ye, Y., Zheng, Y., & Lai, J. (2025). A comprehensive analysis of r's application in ecological research from 2008 to 2023. *Journal of Plant Ecology*, *18*(1), rtaf010. https://doi.org/10.1093/jpe/rtaf010

Kitzes, J., Turek, D., & Deniz, F. (2018). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences.* (J. Kitzes, D. Turek, & F. Deniz, Eds.; pp. 1–319). University of California Press. http://www.practicereproducibleresearch.org

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Part III: Preregistration Template for Ecological Modelling

Adaptive Preregistration for Modelling

A template suitable for use in Ecology and Conservation with accompanying rationale and explanation.

Problem Formulation

> **i** Rationale & Explanation
>
> This section specifies the decision-making context in which the model will be used or the intended scope and context of conclusions. Important components include the decision maker and stakeholders (including experts) and their view on: i) the nature of the problem or decision addressed and how the scope of the modelling tool fits within the (broader) context (i.e. model purpose; ii) the spatial and temporal scales relevant to the decision context; iii) specified desired outputs; iv) role and inclusion in model development and testing; v) whether they foresee unacceptable outcomes that need to be represented in the model (i.e. as constraints), and; vi) what future scenarios does the model need to account for (noting this may be revised later). It should also provide a summary of the domain of applicability of the model, and reasonable extrapolation limits (Grimm et al., 2014).

Model Context and Purpose

> **i** Rationale & Explanation
>
> Defining the purpose of the model is critical because the model purpose influences choices at later stages of model development (Jakeman et al., 2006). Common model purposes in ecology include: gaining a better qualitative understanding of the target system, synthesising and reviewing knowledge, and providing guidance for management and decision-making (Jakeman et al., 2006). Note that modelling objectives are distinct from the analytical objectives of the model.

The scope of the model includes temporal and spatial resolutions, which should also be defined here (Mahmoud et al., 2009). Any external limitations on model development, analysis and flexibility should also be outlined in this section (Jakeman et al., 2006).

Key stakeholders and model users

Preregistration Item

Identify relevant interest groups:

- ☐ Who is the model for?
- ☐ Who is involved in formulating the model?
- ☐ How will key stakeholders be involved in model development?
- ☐ Describe the decision-making context in which the model will be used (if relevant).

Model purpose, context and problem context

Preregistration Item

Briefly outline:

- ☐ the ecological problem,
- ☐ the decision problem (if relevant), including the decision-trigger and any regulatory frameworks relevant to the problem,
- ☐ how the model will address the problem, being clear about the scope of the model i.e. is the model addressing the whole problem, or part of it? Are there any linked problems that your model should consider?
- ☐ Ensure that you specify any focal taxa and study objectives.

Adaptive Preregistration for Modelling

Analytical objectives

Explanation

How will the model be analysed, what analytical questions will the model be used to answer? For example, you might be using your model in a scenario analysis to determine which management decision is associated with minimum regret or the highest likelihood of improvement. Other examples from ecological decision-making include: to compare the performance of alternative management actions under budget constraint (Fraser et al., 2017) to search for robust decisions under uncertainty (McDonald-Madden et al., 2008), to choose the conservation policy that minimises uncertainty [insert ref]. See other examples in (Moallemi et al., 2019).

Preregistration Item

Provide detail on the analytical purpose and scope of the model:

- ☐ How will the model be analysed and what analytical questions will the model be used to answer?
- ☐ Candidate decisions should be investigated and are specified a priori. Depending on the modelling context, they may be specified by stakeholders, model users or the analyst (Moallemi et al., 2019).
  - ☐ Describe the method used to identify relevant management actions and
  - ☐ specify management actions to be considered included in the model.
  - ☐ Are there potentially unacceptable management or policy outcomes identified by stakeholders that should be captured in the model, i.e. as constraints?
- ☐ Are there scenarios that model inputs or outputs that must accommodated? Scenarios should be set a priori, (i.e. before the model is built, Moallemi et al., 2019) and may be stakeholder-defined or driven by the judgement of the modeller or other experts (Mahmoud et al., 2009).
  - ☐ If relevant, describe what processes you will use to elicit and identify relevant scenarios, e.g. literature review, structured workshops with stakeholders or decision-makers.
  - ☐ Specify scenarios under which decisions are investigated.

Adaptive Preregistration for Modelling

Logistical Constraints

Preregistration Item

- ☐ What degree of flexibility is required from the model? Might the model need to be quickly reconfigured to explore new scenarios or problems proposed by clients / managers / model-users?
- ☐ Are there any limitations on model development analysis and flexibility, such as time or budget constraints, for example, does a model need to be deployed rapidly?
    - ☐ When must the model be completed by, e.g. to help make a decision?

Model Scope, Scale and Resolution

Preregistration Item

- ☐ The choice of a model's boundaries is closely linked to the choice of how finely to aggregate the behaviour within the model (Jakeman et al., 2006) - what is the intended scale, and resolution of the model (temporal, spatial or otherwise)?
- ☐ Where is the boundary of the modelled system? Everything outside beyond the boundary and not crossing it is to be ignored within the domain of the model, and everything crossing the boundary is to be treated as external forcing (known/unknown), or else as model outputs (observed, or not, Jakeman et al., 2006).

Intended application of results

Explanation

Preregistration Items in this section are relevant to model transferability (Yates et al., 2018) and constraints on generality in model analysis interpretation. How far do can the results be extrapolated based on the study design (data + model + analysis)? For instance, if there are many confounding variables and not enough spatial / environmental replication, then making broader more general claims beyond the stated boundaries of the model (Section) may not be warranted. However, larger generalisations about results may be acceptable if the data comes from experimentally manipulated or controlled systems.

Adaptive Preregistration for Modelling

Preregistration Item

- ☐ What is the intended domain in which the model is to be applied? Are there any reasonable extrapolation limits beyond which you expect the model should not be applied (Grimm et al., 2014)?

Scenario Analysis Operationalisation

Preregistration Item (delete as necessary)

- ☐ How will you operationalise any scenarios identified in 1.1.3? For example, how will you operationalise any qualitative changes of interest, such as , 'deterioration' or 'improvement'?
- ☐ Describe how you will evaluate and distinguish the performance of alternative scenario outcomes
- ☐ Justify or otherwise explain how you chose these measures and determined performance criteria in relation to the analytical objectives, model purpose and modelling context, such as the risk attitudes of decision-makers and stakeholders within this system

Define Conceptual Model

Explanation

Conceptual models underpin the formal or quantitative model (Cartwright et al., 2016). The conceptual model describes the biological mechanisms relevant to the ecological problem and should capture basic premises about how the target system works, including any prior knowledge and assumptions about system processes. Conceptual models may be represented in a variety of formats, such as influence diagrams, linguistic model block diagram or bond graphs, and these illustrate how model drivers are linked to both outputs or observed responses, and internal (state) variables (Jakeman et al., 2006).

Adaptive Preregistration for Modelling

Choose elicitation and representation method

Preregistration Item

- ☐ Describe what method you will use to elicit or identify the conceptual model. Some common methods include interviews, drawings, and mapping techniques including influence diagrams, cognitive maps and Bayesian belief networks (Moon et al., 2019). It is difficult to decide and justify which method is most appropriate, see Moon et al. (2019) for guidance addressing this methodological question.
- ☐ Finally, how do you intend on representing the final conceptual model? This will likely depend on the method chosen to elicit the conceptual model.

Explain Critical Conceptual Design Decisions

Preregistration Item

List and explain critical conceptual design decisions (Grimm et al., 2014), including:

- ☐ spatial and temporal scales,
- ☐ selection of entities and processes,
- ☐ representation of stochasticity and heterogeneity,
- ☐ consideration of local versus global interactions, environmental drivers, etc.
- ☐ Explain and justify the influence of particular theories, concepts, or earlier models against alternative conceptual design decisions that might lead to alternative model structures.

Model assumptions and uncertainties

Preregistration Item

Specify key assumptions and uncertainties underlying the model design, describing how uncertainty and variation will be represented in the model (Moallemi et al., 2019). Sources of uncertainty may include:

- ☐ exogenous uncertainties affecting the system,
- ☐ parametric uncertainty in input data and
- ☐ structural / conceptual nonparametric uncertainty in the model.

Adaptive Preregistration for Modelling

Identify predictor and response variables

Explanation

The identification and definition of primary model input variables should be driven by scenario definitions, and by the scope of the model described in the problem formulation phase (Mahmoud et al., 2009).

Preregistration Item

Identify and define system system variables and structures, referencing scenario definitions, and the scope of the model as described within problem formulation:

- ☐ What variables would support taking this action or making this decision?
- ☐ What additional variables may interact with this system (things we can't control, but can hopefully measure)?
- ☐ What variables have not been measured, but may interact with the system (often occurs in field or observational studies)?
- ☐ What variables are index or surrogate measures of variables that we cannot or have not measured?
- ☐ In what ways do we expect these variables to interact (model structures)?
- ☐ Explain how any key concepts or terms within problem or decision-making contexts, such as regulatory terms, will be operationalised and defined in a biologically meaningful way to answer the research question appropriately?

Define prior knowledge, data specification and evaluation

Explanation

This section specifies the plan for collecting, processing and preparing data available for parameterisation, determining model structure, and for scenario analysis. It also allows the researchers to disclose any prior interaction with the data.

Adaptive Preregistration for Modelling

Collate available data sources that could be used to parameterise or structure the model

Preregistration Item

For pre-existing data (delete as appropriate):

- ☐ Document the identity, quantity and provenance of any data that will be used to develop, identify and test the model.
- ☐ For each dataset, is the data open or publicly available?
- ☐ How can the data be accessed? Provide a link or contact as appropriate, indicating any restrictions on the use of data.
- ☐ Date of download, access, or expected timing offuture access.
- ☐ Describe the source of the data - what entity originally collected this data? (National Data Set, Private Organisational Data, Own Lab Collection, Other Lab Collection, External Contractor, Meta-Analysis, Expert Elicitation, Other).
- ☐ Codebook and meta-data. If a codebook or other meta-data is available, link to it here and / or upload the document(s).
- ☐ Prior work based on this dataset - Have you published / presented any previous work based on this dataset? Include any publications, conference presentations (papers, posters), or working papers (in-prep, unpublished, preprints) based on this dataset you have worked on.
- ☐ Unpublished Prior Research Activity - Describe any prior but unpublished research activity using these data. Be specific and transparent.
- ☐ Prior knowledge of the current dataset - Describe any prior knowledge of or interaction with the dataset before commencing this study. For example, have you read any reports or publications about this data?
- ☐ Describe how the data is arranged, in terms of replicates and covariates.

Sampling Plan (for data you will collect, delete as appropriate):

- ☐ Data collection procedures - Please describe your data collection process, including how sites and transects or any other physical unit were selected and arranged. Describe any inclusion or exclusion rules, and the study timeline.
- ☐ Sample Size - Describe the sample size of your study.
- ☐ Sample Size Rationale - Describe how you determined the appropriate sample size for your study. It could include feasibility constraints, such as time, money or personnel.
- ☐ If sample size cannot be specified, specify a stopping rule - i.e. how will you decide when to terminate your data collection?

Adaptive Preregistration for Modelling

Data Processing and Preparation

Preregistration Item

- ☐ Describe any data preparation and processing steps, including manipulation of environmental layers (e.g. standardisation and geographic projection) or variable construction (e.g. Principal Component Analysis).

Describe any data exploration or preliminary data analyses.

Explanation

In most modelling cases, it is necessary to perform preliminary analyses to understand the data and check that assumptions and requirements of the chosen modelling procedures are met. Data exploration prior to model fitting or development may include exploratory analyses to check for collinearity, spatial and temporal coverage, quality and resolution, outliers, or the need for transformations (Yates et al., 2018).

Preregistration Item

For each separate preliminary or investigatory analysis:

- ☐ State what needs to be known to proceed with further decision-making about the modelling procedure, and why the analysis is necessary.
- ☐ Explain how you will implement this analysis, as well as any techniques you will use to summarise and explore your data.
- ☐ What method will you use to represent this analysis (graphical, tabular, or otherwise, describe)
- ☐ Specify exactly which parts of the data will be used
- ☐ Describe how the results will be interpreted, listing each potential analytic decision, as well as the analysis finding that will trigger each decision, where possible.

Adaptive Preregistration for Modelling

Data evaluation, exclusion and missing data

Explanation

Documenting issues with reliability is important because data quality and ecological relevance might be constrained by measurement error, inappropriate experimental design, and heterogeneity and variability inherent in ecological systems (Grimm et al., 2014). Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand (Mahmoud et al., 2009).

Preregistration Item

- ☐ Describe how you will determine how reliable the data is for the given model purpose. Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand
- ☐ Document any issues with data reliability.
- ☐ How will you determine what data, if any, will be excluded from your analyses?
- ☐ How will outliers be handled? Describe rules for identifying outlier data, and for excluding a site, transect, quadrat, year or season, species, trait, etc.
- ☐ How will you identify and deal with incomplete or missing data?

Conceptual model evaluation

Preregistration Item

- ☐ Describe how your conceptual model will be critically evaluated. Evaluation includes both the completeness and suitability of the overall model structure.
- ☐ How will you critically assess any simplifying assumptions (Augusiak et al., 2014)?
- ☐ Will this process will include consultation or feedback from a client, manager, or model user.

Adaptive Preregistration for Modelling

Formalise and Specify Model

Explanation

In this section describe what quantitative methods you will use to build the model/s, explain how they are relevant to the client/manager/user's purpose.

Model class, modelling framework and approach

Explanation

Modelling approaches can be described as occurring on a spectrum from correlative or phenomenological to mechanistic or process-based (Yates et al., 2018); where correlative models use mathematical functions fitted to data to describe underlying processes, and mechanistic models explicitly represent processes and details of component parts of a biological system that are expected to give rise to the data (White & Marshall, 2019). A model 'class,' 'family' ' or 'type' is often used to describe a set of models each of which has a distinct but related sampling distribution (C. C. Liu & Aitkin, 2008). The model family is driven by choices about the types of variables covered and the nature of their treatment, as well as structural features of the model, such as link functions, spatial and temporal scales of processes and their interactions (Jakeman et al., 2006).

Preregistration Item

- ☐ Describe what modelling framework, approach or class of model you will use to implement your model and relate your choice to the model purpose and analytical objectives described in 1.1.2 and 1.1.3.

Choose model features and family

Explanation

All modelling approaches require the selection of model features, which conform with the conceptual model and data specified in previous steps (Jakeman et al., 2006). The choice of model are determined in conjunction with features are selected. Model features include elements such as the functional form of interactions, data structures, measures used to specify links, any bins or discretisation of continuous variables. It is usually difficult to change fundamental features of a model beyond

an early stage of model development, so careful thought and planning here is useful to the modeller (Jakeman et al., 2006). However, if changes to these fundamental aspects of the model do need to change, document how and why these choices were made, including any results used to support any changes in the model.

Operationalising Model Variables

Preregistration Item

- ☐ For each response, predictor, and covariate, *specify how these variables will be operationalised in the model. This should relate directly to the analytical and/or management objectives specified during the problem formulation phase. Operationalisations could include: the extent of a response, an extreme value, a trend, a long-term mean, a probability distribution, a spatial pattern, a time-series, qualitative change, such as a direction of change or, the frequency, location, or probability of some event occuring. Specify any treatment of model variables, including whether they are lumped / distributed, lienar / non-linear, stochastic / deterministic (Jakeman et al., 2006)* .
- ☐ *Provide a rationale for your choices, including why plausible alternatives under consideration were not chosen, and relate your justification bacj to the purpose, objectives, prior knowledge and or logistical constraints specified in the problem formulation phase (Jakeman et al., 2006).*

Choose model family

Preregistration Item

- ☐ Specify which family of statistical distributions you will use in your model, and describe any transformations, or link functions.
- ☐ Include in your rational for selection, detail about which variables the model outputs are likely sensitive to, what aspects of their behaviour are important, and any associated spatial or temporal dimensions in sampling.

Adaptive Preregistration for Modelling

Describe *approach* for identifying model structure

Explanation

This section relates to the process of determining the best/most efficient/parsimonious representation of the system at the appropriate scale of concern (Jakeman et al., 2006) that best meets the analytical objectives specified in the problem formulation phase. Model structure refers to the choice of variables included in the model, and the nature of the relationship among those variables. Approaches to finding model structure and parameters may be knowledge-supported, or data-driven (Boets et al., 2015). Model selection methods can include traditional inferential approaches such as unconstrained searches of a dataset for patterns that explain variations in the response variable, or use of ensemble-modelling methods (Barnard et al., 2019). Ensemble modelling procedures might aim to derive a single model, or a multi-model average (Yates et al., 2018). Refining actions to develop a model could include iteratively dropping parameters or adding them, or aggregating / disaggregating system descriptors, such as dimensionality and processes (Jakeman et al., 2006).

Preregistration Item

- ☐ Specify what approach and methods you will use to identify model structure and parameters.
- ☐ If using a knowledge-supported approach to deriving model structure (either in whole or in part), specify model structural features, including:
  - the functional form of interactions (if any)
  - data structures,
  - measures used to specify links,
  - any bins or discretisation of continuous variables (Jakeman et al., 2006),
  - any other relevant features of the model structure.

Describe parameter estimation technique and performance criteria

Explanation

Before calibrating the model to the data, the performance criteria for judging the calibration (or model fit) are specified. These criteria and their underlying assumptions should reflect the desired properties of the parameter estimates / structure (Jakeman et al., 2006). For example, modellers might seek parameter estimates that are robust to outliers, unbiased, and yield appropriate predictive

performance. Modellers will need to consider whether the assumptions of the estimation technique yielding those desired properties are suited to the problem at hand. For integrated or sub-divided models, other considerations might include choices about where to disaggregate the model for parameter estimation; e.g. spatial sectioning (streams into reaches) and temporal sectioning (piece-wise linear models) (Jakeman et al., 2006).

Parameter estimation technique

Preregistration Item

- ☐ Specify what technique you will use to estimate parameter values, and how you will supply non-parametric variables and/or data (e.g. distributed boundary conditions). For example, will you calibrate all variables simultaneously by optimising fit of model outputs to observations, or will you parameterise the model in a piecemeal fashion by either direct measurement, inference from secondary data, or some combination (Jakeman et al., 2006).
- ☐ Identify which variables will be parameterised directly, such as by expert elicitation or prior knowledge.
- ☐ Specify which algorithm(s) you will use for any data-driven parameter estimation, including supervised, or unsupervised machine learning, decision-tree, K-nearest neighbour or cluster algorithms (Z. Liu et al., 2018).

Parameter estimation / model fit performance criteria

Preregistration Item

- ☐ Specify which suite of performance criteria you will use to judge the performance of the model. Examples include correlation scores, coefficient of determination, specificity, sensitivity, AUC, etcetera (Yates et al., 2018).
- ☐ Relate any underlying assumptions of each criterion to the desired properties of the model, and justify the choice of performance metric in relation
- ☐ Explain how you will identify which model features or components are significant or meaningful.

Adaptive Preregistration for Modelling

Model assumptions and uncertainties

Preregistration Item

- ☐ Specify assumptions and key uncertainties in the formal model. Describe what gaps exist between the model conception, and the real-world problem, what biases might this introduce and how might this impact any interpretation of the model outputs, and what implications are there for evaluating model-output to inform inferences or decisions?

Specify formal model(s)

Explanation

Once critical decisions have been made about the modelling approach and method of model specification, the conceptual model is translated into the quantitative model.

Preregistration Item

- ☐ Specify all formal models
  - ☐ Note, For data-driven approaches to determining model structure and or parameterisation, it may not be possible to respond to this preregistration item. In such cases, explain why this is the case, and how you will document the model(s) used in the final analysis.
- ☐ For quantitative model selection approaches, including ensemble modelling, specify each model used in the candidate set, including any null or full/global model.

Adaptive Preregistration for Modelling

Model Calibration, Validation & Checking

Model calibration and validation scheme

Explanation

This section pertains to any data calibration, validation or testing schemes that will be implemented. For example, the model may be tested on data independent of those used to parameterise the model (external validation), or the model may be cross-validated on random sub-samples of the data used to parameterise the model (internal cross-validation) (Barnard et al., 2019; Yates et al., 2018). For some types of models, hyper-parameters are estimated from data, and may be tuned on further independent holdouts of the training data, ("validation data").

Preregistration Item

- ☐ Describe any data calibration, validation and testing scheme you will implement, including any procedures for tuning or estimating model hyper-parameters (if any).

Describe calibration/validation data

Explanation & Rationale

The following items pertain to properties of the *datasets* used for calibration (training), validation, and testing.

Preregistration Item

If partitioning data for cross-validation or similar approach (delete as needed):

- ☐ Describe the approach specifying the number of folds that will be created, the relative size of each fold, and any stratification methods used for ensuring evenness of groups between folds and between calibration / validation data?

If using external / independent holdout data for model testing and evaluation (delete as needed):

- ☐ Which data will be used as a the testing data? What method will you be used for generating training / test data subsets?

- ☐ Describe any known differences between the training/validation and testing datasets, the relative size of each, as well as any stratification methods used for ensuring evenness of groups between data sets?
- ☐ It is preferable that any independent data used for model testing remains unknown to modellers during the process of model development, please describe the relationship modellers have to model validation data, will independent datasets be known or accessible to any modeller or analyst?

Implementation verification

Explanation & Examples

Model implementation verification is the process of ensuring that the model has been correctly implemented, and that the model performs as described by the model description (Grimm et al., 2014). This process is distinct from model checking, which assesses the model's performance in representing the system of interest (Conn et al., 2018).

- Checks for verification implementation should include i) thoroughly checking for bugs or programming errors, and ii) whether the implemented model performs as described by the model description (Grimm et al., 2014).
- Qualitative tests could include syntax checking of code, and peer-code review (Ivimey et al., 2023). Technical measures include using unit tests, or in-built checks within functions to prevent potential errors.

Preregistration Item

- ☐ What Quality Assurance measures will you take to verify the model has been correctly implemented? Specifying a priori quality assurance tests for implementation verification may help to avoid selective debugging and silent errors.

Adaptive Preregistration for Modelling

Model checking

Rationale & Explanation

"Model Checking" goes by many names ("conditional verification", "quantitative verification", "model output verification"), and refers to a series of analyses that assess a model's performance in representing the system of interest (Conn et al., 2018). Model checking aids in diagnosing assumption violations, and reveals where a model might need to be altered to better represent the data, and therefore system (Conn et al., 2018). Quantitative model checking diagnostics include goodness of fit, tests on residuals or errors, such as for heteroscedascity, cross-correlation, and autocorrelation (Jakeman et al., 2006).

Quantitative model checking

Preregistration Item

During this process, observed data, or data and patterns that guided model design and calibration, are compared to model output in order to identify if and where there are any systematic differences.

- ☐ Specify any diagnostics or tests you will use during model checking to assess a model's performance in representing the system of interest.
- ☐ For each test, specify the criteria that will you use to interpret the outcome of the test in assessing the model's ability to sufficiently represent the gathered data used to develop and parameterise the model.

Qualitative model checking

Explanation

This step is largely informal and case-specific, but requires, 'face validation' with model users / clients / managers who aren't involved in the development of the model to assess whether the interactions and outcomes of the model are feasible an defensible (Grimm et al., 2014). This process is sometimes called a ,"laugh test" or a "pub test" and in addition to checking the model's believability, it builds the client's confidence in the model (Jakeman et al., 2006). Face validation could include structured walk-throughs, or presenting descriptions, visualisations or summaries of model results to experts for assessment.

Adaptive Preregistration for Modelling

Preregistration Item

☐ Briefly explain how you will qualitatively check the model, and whether and how you will include users and clients in the process.

Assumption Violation Checks

Preregistration Item

The consequences of assumption violations on the interpretation of results should be assessed (Araújo et al., 2019).

☐ Explain how you will demonstrate robustness to model assumptions and check for violations of model assumptions.
☐ If you cannot perform quantitative assumption checks, describe what theoretical justifications would justify a lack of violation of or robustness to model assumptions.
☐ If you cannot demonstrate or theoretically justify violation or robustness to assumptions, explain why not, and specify whether you will discuss assumption violations and their consequences for interpretation of model outputs.
☐ If assumption violations cannot be avoided, explain how you will explore the consequences of assumption violations on the interpretation of results (To be completed in interim iterations of the preregistration, only if there are departures from assumptions as demonstrated in the planned tests above).

Model Validation and Evaluation

Explanation

The model validation & evaluation phase comprises a suite of analyses that collectively inform inferences about whether, and under what conditions, a model is suitable to meet its intended purpose (Augusiak et al., 2014). Errors in design and implementation of the model and their implication on the model output are assessed. Ideally independent data is used against the model outputs to assess whether the model output behaviour exhibits the required accuracy for the model's intended purpose. The outcomes of these analyses build confidence in the model applications and increase understanding of model strengths and limitations. Model evaluation including, model analysis, should complement model checking. It should evaluate model checking, and consider over-fitting and extrapolation. The higher the propor-

tion of calibrated, or uncertain parameters, "the greater the risk that the model seems to work correctly, but for the wrong reasons" (citaiton). Evaluation thus complements model checking because we can rule out the chance that the model fits the calibration data well, but has not captured the relevant ecological mechanisms of the system pertinent to the research question or the decision problem underpinning the model (Grimm et al., 2014). Evaluation of model outputs against external data in conjunction with the results from model checking provide information about the structural realism and therefore credibility of the model (Grimm & Berger, 2016).

Model output corroboration

Explanation

Ideally, model outputs or predictions are compared to independent data and patterns that were not used to develop, parameterise, or verify the model. Testing against a dataset of response and predictor variables that are spatially and/or temporally independent from the training dataset minimises the risk of artificially inflating model performance measures (Araújo et al., 2019). Although the corroboration of model outputs against an independent validation dataset is considered the 'gold standard' for showing that a model properly represents the internal organisation of the system, model validation is not always possible because empirical experiments are infeasible or model users are working on rapid-response time-frames, hence, why ecologists often model in the first place (Grimm et al., 2014). Independent predictions might instead be tested on sub-models. Alternatively, patterns in model output that are robust and seem characteristic of the system can be identified and evaluated in consultation with the literature or by experts to judge how accurate the model output is (Grimm et al., 2014).

Preregistration Item

- ☐ State whether you will corroborate the model outputs on external data, and document any independent validation data in step.
- ☐ It is preferable that any independent data used for model evaluation remains unknown to modellers during the process of model building (Dwork et al., 2015), describe the relationship modellers have to model validation data, e.g. will independent datasets be known to any modeller or analyst involved in the model building process?
- ☐ If unable to evaluate the model outputs against independent data, explain why and explain what steps you will take to interrogate the model.

Choose performance metrics and criteria

Explanation

Model performance can be quantified by a range of tests, including measures of agreement between predictions and independent observations, or estimates of accuracy, bias, calibration, discrimination refinement, resolution and skill (Araújo et al., 2019). Note that the performance metrics and criteria in this section are used for evaluating the structured and parameterised models (ideally) on independent holdout data, so this step is additional to any performance criteria used for determining model structure or parameterisation in section X.

Preregistration Item

- ☐ Specify what performance measures you will use to evaluate the model and briefly explain how each test relates to different desired properties of a model's performance.
- ☐ Spatial, temporal and environmental pattern of errors and variance can change the interpretation of model predictions and conservation decisions (Araújo et al., 2019), where relevant and possible, describe how you will characterise and report the spatial, temporal and environmental pattern of errors and variance.
- ☐ If comparing alternative models, specify what measures of model comparison or out-of-sample performance metrics will you use to find support for alternative models or else to optimise predictive ability. State what numerical threshold or qualities you will use for each of these metrics.

Model analysis

Rationale & Explanation

Uncertainty in models arises due to incomplete system understanding (which processes to include, or which interact), from imprecise, finite and sparese data measurements, and from uncertainty in input conditions and scenarios for model simulations or runs (Jakeman et al., 2006). Non-technical uncertainties can also be introduced throughout the modellign process, such as uncertainties arising from issues in problem-framing, indeterminicies, and modeller / client values (Jakeman et al., 2006).
The purpose of model analysis is to prevent blind trust in the model by understanding how model outputs have emerged, and to 'challenge' the model by verifying whether the model is still believable and fit for purpose if one or more parameters are changed

(Grimm et al., 2014).
Model analysis should increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system. Model analysis typically consists of sensitivity analyses preceded by uncertainty analyses (Saltelli et al., 2019), and a suite of other simulation or other computational experiments. The aim of such computational experiments is to increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system (Grimm et al., 2014). Uncertainty analyses and sensitivity analyses augment one another to draw conclusions about model uncertainty.
Because the results from a full suite of sensitivity analysis and uncertainty analysis can be difficult to interpret due to the number and complexity of causal relations examined (Jakeman et al., 2006), it is useful for the analyst to relate the choice of analysis to the modelling context, purpose and analytical objectives defined in the problem formulation phase, in tandem with any critical uncertainties that have emerged during model development and testing prior to this point.

Uncertainty Analyses

Explanation

Uncertainty can arise from different modelling techniques, response data and predictor variables (Araújo et al., 2019). Uncertainty analyses characterise the uncertainty in model outputs, and identify how uncertainty in model parameters affects uncertainty in model output, but does not identify which model assumptions are driving this behaviour (Grimm et al., 2014; Saltelli et al., 2019). Uncertainty analyses can include propagating known uncertainties through the model, or by investigating the effect of different model scenarios with different parameters and modelling technique combinations (Araújo et al., 2019), for example. It could also include characterising the output distribution, such as through empirical construction using model output data points. It could also include extracting summary statistics like the mean, median and variance from this distribution, and perhaps constructing confidence intervals on the mean (Saltelli et al., 2019).

Preregistration Item

☐ Please describe how you will characterise model and data uncertainties, e.g. propagating known uncertainties through the model, investigating the effect of different model scenarios with different parameters and modelling technique combinations (Araújo et al., 2019), or empirically constructing

model distributions from model output data points, and extracting summary statistics, including the mean, median, variance, and constructing confidence intervals (Saltelli et al., 2019).

- ☐ Relate your choice of analysis to the context and purposes of the model described in the problem formulation phase. For instance , discrepancies between model output and observed output may be important for forecasting models, where cost, benefit, an risk over a substantial period must be gauged, but much less critical for decision-making or management models where the user may be satisfied with knowing that the predicted ranking order of impacts of alternative scenarios or management options is likely to be correct, with only a rough indication of their sizes" (Jakeman et al., 2006).
- ☐ Briefly describe how you will summarise the results of these in silico experiments with graphical, tabular, or other devices, such as summary statistics.
- ☐ If the chosen modelling approach is able to explicitly articulate uncertainty due to data, measurements or baseline conditions, such as by providing estimates of uncertainty (typically in the form of probabilistic parameter covariance, (Jakeman et al., 2006)), specify which measure of uncertainty you will use.

Sensitivity analyses

Explanation

Sensitivity analysis examines how uncertainty in model outputs can be apportioned to different sources of uncertainty in model input (Saltelli et al., 2019).

Preregistration Item

- ☐ Describe the sensitivity analysis approach you will take: deterministic sensitivity, stochastic sensitivity (variability in the model), or scenario sensitivity (effect of changes based on scenarios).
- ☐ Describe any sensitivity analyses you will conduct by specifying which parameters will be held constant, which will be varied, and the range and intervals of values over which those parameters will be varied.
- ☐ State the primary objective of each sensitivity analysis, for example, to identify which input variables contribute the most to model uncertainty so that these variables can be targeted for further data collection, or alternatively to identify which variables or factors contribute little to overall model outputs, and so can be 'dropped' from future iterations of the model (Saltelli et al., 2019).

Model application or scenario analysis

Preregistration Item

- ☐ Specify any input conditions and relevant parameter values for initial environmental conditions and decision-variables under each scenario specified in section 1.
- ☐ Describe any other relevant technical details of model application, such as methods for how you will implement any simulations or model projections.
- ☐ What raw and transformed model outputs will you extract from the model simulations or projections, and how will you map, plot, or otherwise display and synthesise the results of scenario and model analyses.
- ☐ Explain how you will analyse the outputs to answer your analytical objectives. For instance, describe any trade-off or robustness analyses you will undertake to help evaluate and choose between different alternatives in consultation with experts or decision-makers.

Other simulation experiments / robustness analyses

Preregistration Item

- ☐ Describe any other simulation experiments, robustness analyses or other analyses you will perform on the model, including any metrics and their criteria / thresholds for interpreting the results of the analysis.

References

Araújo, M., Anderson, R., Márcia Barbosa, A., Beale, C., Dormann, C., Early, R., Garcia, R., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R., Zimmermann, N., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Barnard, D. M., Germino, M. J., Pilliod, D. S., Arkle, R. S., Applestein, C., Davidson, B. E., & Fisk, M. R. (2019). Can't see the random forest for the decision trees: Selecting predictive models for restoration ecology. *Restoration Ecology*, *27*(5), 1053–1063. https://doi.org/10.1111/rec.12938

Boets, P., Landuyt, D., Everaert, G., Broekx, S., & Goethals, P. L. M. (2015). *Evaluation and comparison of data-driven and knowledge-supported bayesian belief networks to assess the habitat suitability for alien macroinvertebrates. 74*, 92–103. https://doi.org/10.1016/j.envsoft.2015.09.005

Cartwright, S. J., Bowgen, K. M., Collop, C., Hyder, K., Nabe-Nielsen, J., Stafford, R., Stillman, R. A., Thorpe, R. B., & Sibly, R. M. (2016). Communicating complex ecological models to non-scientist end users. *Ecological Modelling*, *338*, 51–59. https://doi.org/10.1016/j.ecolmodel.2016.07.012

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *9*, 341–317. https://doi.org/10.1002/ecm.1314

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Fraser, H., Rumpff, L., Yen, J. D. L., Robinson, D., & Wintle, B. A. (2017). Integrated models to support multiobjective ecological restoration decisions. *Conservation Biology*, *31*(6), 1418–1427. https://doi.org/10.1111/cobi.12939

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129–139. https://doi.org/0.1016/j.ecolmodel.2014.01.018

Grimm, V., & Berger, U. (2016). Structural realism, emergence, and predictions in next-generation ecological modelling: Synthesis from a special issue. *Ecological Modelling*, *326*, 177–187. https://doi.org/10.1016/j.ecolmodel.2016.01.001

Ivimey, E. R., Pick, J. L., Bairos, K. R., Culina, A., Gould, E., Grainger, M., Marshall, B. M., Moreau, D., Paquet, M., Royauté, R., Sánchez, A., Silva, I., & Windecker, S. M. (2023). Implementing code review in the scientific workflow: Insights from ecology and evolutionary biology. *Journal of Evolutionary Biology*, *36*, 1347–1356. https://doi.org/10.1111/jeb.14230

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., & Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: Recent progress and future challenges. *Environmental Reviews*, *26*(4), 339–350. https://doi.org/10.1139/er-2018-0034

Mahmoud, M., Liu, Y., Hartmann, H., Stewart, S., Wagener, T., Semmens, D., Stewart, R., Gupta, H., Dominguez, D., Dominguez, F., Hulse, D., Letcher, R., Rashleigh, B., Smith, C., Street, R., Ticehurst, J., Twery, M., Delden, H. van, Waldick, R.,

… Winter, L. (2009). A formal framework for scenario development in support of environmental decision-making. *Environmental Modelling & Software*, *24*(7), 798–808. https://doi.org/10.1016/j.envsoft.2008.11.010

McDonald-Madden, E., Baxter, P. W. J., & Possingham, H. P. (2008). Making robust decisions for conservation with restricted money and knowledge. *Journal of Applied Ecology*, *45*(6), 1630–1638. https://doi.org/10.1111/j.1365-2664.2008.01553.x

Moallemi, E. A., Elsawah, S., & Ryan, M. J. (2019). Strengthening "good" modelling practices in robust decision support: A reporting guideline for combining multiple model-based methods. *Mathematics and Computers in Simulation*, *175*, 3–24. https://doi.org/10.1016/j.matcom.2019.05.002

Moon, K., Guerrero, A. M., Adams, Vanessa. M., Biggs, D., Blackman, D. A., Craven, L., Dickinson, H., & Ross, H. (2019). Mental models for conservation research and practice. *Conservation Letters*, *12*(3), e12642. https://doi.org/10.1111/conl.12642

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29–39. https://doi.org/10.1016/j.envsoft.2019.01.012

White, C. R., & Marshall, D. J. (2019). Should we care if models are phenomenological or mechanistic. *Trends in Ecology & Evolution*, *34*(4), 276–278. https://doi.org/10.1016/j.tree.2019.01.006

Yates, K., Bouchet, P., Caley, M., Mengersen, K., Randin, C., Parnell, S., Fielding, A., Bamford, A., Ban, S., Barbosa, A., Dormann, C., Elith, J., Embling, C., Ervin, G., Fisher, R., Gould, S., Graf, R., Gregr, E., Halpin, P., … Sequeira, A. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

D.4 Ecological Modelling Preregistration Template

The preregistration template is reproduced below. The template is archived on Zenodo at:

> Gould, E. (2025). EcoConsModPreReg: A Quarto Template for Preregistering Ecological Modelling Studies (v.0.0.9005). Zenodo. https://doi.org/10.5281/zenodo.15655052

And is available for installation as a quarto template from github.com/egouldo/EcoConsMod-PreReg?tab=readme-ov-file, where the `.docx` template format can also be downloaded.

Ecological Modelling Preregistration Template

Elliot Gould[1], Christopher S. Jones[12], Jian D.L. Yen[2], Hannah S. Fraser[1], Henry Wootton[2], Megan K. Good[1], David H. Duncan[1], Cindy E. Hauser[2], Bonnie C. Wintle[1], Libby Rumpff[1]

[1]School of Agriculture, Food and Ecosystem Sciences, University of Melbourne
[2]Arthur Rylah Institute for Environmental Research, Department of Energy, Environment and Climate Action

2026-04-07

Background and Instructions

Here we present a preregistration template for ecological models in ecology, conservation and related fields.

Replace author, author-affiliations and persistent ID's (e.g. ORCID iD), keywords, title and abstract metadata in the document YAML as relevant to your study.

For non-trivial modelling studies, especially where model parameter and structure is in any way data-contingent, we recommend taking an *Adaptive Preregistration* approach (Gould, Jones, et al., 2026). Please read the *Primer on Adaptive Preregistration* before beginning.

All preregistration items should be completed, excluding items marked as optional or in cases where they are not applicable to your study, which can be deleted. Additional preregistration items can be added as required at the researchers' discretion.

Table of contents

Study Information **2**
CRediT Contribution Statement . . . 2
Conflict of Interest Statement . . . 2
Data Availability Statement . . . 2
Code Availability . . . 2
Ethics . . . 3

1 Problem Formulation **3**
1.1 Model context and purpose . . . 3
1.2 Scenario analysis operationalisation . . . 5

2 Define Conceptual Model **5**
2.1 Choose elicitation and representation method . . . 6
2.2 Explain critical conceptual design decisions . . . 6
2.3 Model assumptions and uncertainties . . . 6
2.4 Identify predictor and response variables . . . 6
2.5 Define prior knowledge, data specification and evaluation . . . 7
2.6 Conceptual model evaluation . . . 9

3 Formalise and Specify Model **9**
3.1 Model class, modelling framework and approach . . . 9
3.2 Choose model features and family . . . 9
3.3 Describe *approach* for identifying model structure . . . 10
3.4 Describe parameter estimation technique and performance criteria . . . 11
3.5 Model assumptions and uncertainties . . . 11
3.6 Specify formal model(s) . . . 11

4 Model Calibration, Validation & Checking **12**
4.1 Model calibration and validation scheme . . . 12
4.2 Implementation verification . . . 13
4.3 Model checking . . . 13

5 Model Validation and Evaluation **14**

5.1 Model output corroboration 14
5.2 Choose performance metrics and criteria 15
5.3 Model analysis 15

References **17**

Study Information

CRediT Contribution Statement

Preregistration Item

Identify potential contributions according to the CRediT taxonomy (https://doi.org/10.1371/journal.pone.0244611.t001) and write a CRediT contribution statement.

Conflict of Interest Statement

Preregistration Item

☐ Explain any real or perceived conflicts of interest with this study execution. For example, any interests or activities that might be seen as influencing the research (e.g., financial interests in a test or procedure, funding by companies for research).

Data Availability Statement

Preregistration Item

Select one option from below:

☐ "We plan to make the data available (yes / no)," specify the planned data availability level from the following options:
- Data access via download; usage of data for all purposes (public use file)
- Data access via download; usage of data restricted to scientific purposes (scientific use file)
- Data access via download; usage of data has to be agreed and defined on an individual case basis
- Data access via secure data centre (no download, usage/analysis only in a secure data centre)
- Data available upon email request by member of scientific community
- Other (please specify)

☐ "Data will not be made available"
- ☐ Justify reason for not making data available.

Code Availability

Preregistration Item

Select one option from below:

☐ "We plan to make the code available (yes / no)," specify the planned code availability level from the following options:
- Code access via download; usage of code for all purposes (public use file)
- Code access via download; usage of code restricted to scientific purposes (scientific use file)
- Code access via download; usage of code has to be agreed and defined on an individual case basis
- Code access via secure code centre (no download, usage/analysis only in a secure code centre)
- Code available upon email request by member of scientific community
- Other (please specify)

☐ "Code will not be made available"

☐ Justify reason for not making code available.

Ethics

Preregistration Item

☐ Select and respond to the relevant item below:
- If relevant institutional ethical approval for the study has been obtained, provide the relevant identifier, and link to relevant documents.
- If ethical approval has not yet been obtained, but is required, provide a brief overview of plans for obtaining study approval in accordance with established ethical guidelines.
- Alternatively, if the study is exempt from ethical approval, explain exemption.

1 Problem Formulation

Rationale & Explanation

This section specifies the decision-making context in which the model will be used or the intended scope and context of conclusions. Important components include the decision maker and stakeholders (including experts) and their view on: i) the nature of the problem or decision addressed and how the scope of the modelling tool fits within the (broader) context (i.e. model purpose; ii) the spatial and temporal scales relevant to the decision context; iii) specified desired outputs; iv) role and inclusion in model development and testing; v) whether they foresee unacceptable outcomes that need to be represented in the model (i.e. as constraints), and; vi) what future scenarios does the model need to account for (noting this may be revised later). It should also provide a summary of the domain of applicability of the model, and reasonable extrapolation limits (Grimm et al., 2014).

1.1 Model context and purpose

Rationale & Explanation

Defining the purpose of the model is critical because the model purpose influences choices at later stages of model development (Jakeman et al., 2006). Common model purposes in ecology include: gaining a better qualitative understanding of the target system, synthesising and reviewing knowledge, and providing guidance for management and decision-making (Jakeman et al., 2006). Note that modelling objectives are distinct from the analytical objectives of the model.
The scope of the model includes temporal and spatial resolutions, which should also be defined here (Mahmoud et al., 2009). Any external limitations on model development, analysis and flexibility should also be outlined in this section (Jakeman et al., 2006).

1.1.1 Key stakeholders and model users

Preregistration Item

Identify relevant interest groups:
☐ Who is the model for?
☐ Who is involved in formulating the model?
☐ How will key stakeholders be involved in model development?
☐ Describe the decision-making context in which the model will be used (if relevant).

1.1.2 Model purpose, context and problem context

Preregistration Item

Briefly outline:

- ☐ the ecological problem,
- ☐ the decision problem (if relevant), including the decision-trigger and any regulatory frameworks relevant to the problem,
- ☐ how the model will address the problem, being clear about the scope of the model i.e. is the model addressing the whole problem, or part of it? Are there any linked problems that your model should consider?
- ☐ Ensure that you specify any focal taxa and study objectives.

1.1.3 Analytical objectives

Explanation

How will the model be analysed, what analytical questions will the model be used to answer? For example, you might be using your model in a scenario analysis to determine which management decision is associated with minimum regret or the highest likelihood of improvement. Other examples from ecological decision-making include: to compare the performance of alternative management actions under budget constraint (Fraser et al., 2017), to search for robust decisions under uncertainty (McDonald-Madden et al., 2008), to choose the conservation policy that minimises uncertainty (McCarthy et al., 2011). See other examples in (Moallemi et al., 2019).

Preregistration Item

Provide detail on the analytical purpose and scope of the model:

- ☐ How will the model be analysed and what analytical questions will the model be used to answer?
- ☐ Candidate decisions should be investigated and are specified a priori. Depending on the modelling context, they may be specified by stakeholders, model users or the analyst (Moallemi et al., 2019).
  - ☐ Describe the method used to identify relevant management actions and
  - ☐ specify management actions to be considered included in the model.
  - ☐ Are there potentially unacceptable management or policy outcomes identified by stakeholders that should be captured in the model, i.e. as constraints?
- ☐ Are there scenarios that model inputs or outputs that must accommodated? Scenarios should be set a priori, (i.e. before the model is built, Moallemi et al., 2019) and may be stakeholder-defined or driven by the judgement of the modeller or other experts (Mahmoud et al., 2009).
  - ☐ If relevant, describe what processes you will use to elicit and identify relevant scenarios, e.g. literature review, structured workshops with stakeholders or decision-makers.
  - ☐ Specify scenarios under which decisions are investigated.

1.1.4 Logistical Constraints

Preregistration Item

- ☐ What degree of flexibility is required from the model? Might the model need to be quickly reconfigured to explore new scenarios or problems proposed by clients / managers / model-users?
- ☐ Are there any limitations on model development analysis and flexibility, such as time or budget constraints, for example, does a model need to be deployed rapidly?
  - ☐ When must the model be completed by, e.g. to help make a decision?

1.1.5 Model Scope, Scale and Resolution

Preregistration Item

- ☐ The choice of a model's boundaries is closely linked to the choice of how finely to aggregate the behaviour within the model (Jakeman et al., 2006) - what is the intended scale, and resolution of the model (temporal, spatial or otherwise)?
- ☐ Where is the boundary of the modelled system? Everything outside beyond the boundary and not crossing it is to be ignored within the domain of the model, and everything crossing the boundary is to be treated as external forcing (known/unknown), or else as model outputs (observed, or not, Jakeman et al., 2006).

1.1.6 Intended application of results

Explanation

Preregistration Items in this section are relevant to model transferability (Yates et al., 2018) and constraints on generality in model analysis interpretation. How far do can the results be extrapolated based on the study design (data + model + analysis)? For instance, if there are many confounding variables and not enough spatial / environmental replication, then making broader more general claims beyond the stated boundaries of the model (Section 1.1.3) may not be warranted. However, larger generalisations about results may be acceptable if the data comes from experimentally manipulated or controlled systems.

Preregistration Item

- ☐ What is the intended domain in which the model is to be applied? Are there any reasonable extrapolation limits beyond which you expect the model should not be applied (Grimm et al., 2014)?

1.2 Scenario analysis operationalisation

Preregistration Item (delete as necessary)

- ☐ How will you operationalise any scenarios identified in Section 1.1.3? For example, how will you operationalise any qualitative changes of interest, such as , 'deterioration' or 'improvement'?
- ☐ Describe how you will evaluate and distinguish the performance of alternative scenario outcomes
- ☐ Justify or otherwise explain how you chose these measures and determined performance criteria in relation to the analytical objectives, model purpose and modelling context, such as the risk attitudes of decision-makers and stakeholders within this system

2 Define Conceptual Model

Explanation

Conceptual models underpin the formal or quantitative model (Cartwright et al., 2016). The conceptual model describes the biological mechanisms relevant to the ecological problem and should capture basic premises about how the target system works, including any prior knowledge and assumptions about system processes. Conceptual models may be represented in a variety of formats, such as influence diagrams, linguistic model block diagram or bond graphs, and these illustrate how model drivers are linked to both outputs or observed responses, and internal (state) variables (Jakeman et al., 2006).

2.1 Choose elicitation and representation method

Preregistration Item

- ☐ Describe what method you will use to elicit or identify the conceptual model. Some common methods include interviews, drawings, and mapping techniques including influence diagrams, cognitive maps and Bayesian belief networks (Moon et al., 2019). It is difficult to decide and justify which method is most appropriate, see Moon et al. (2019) for guidance addressing this methodological question.
- ☐ Finally, how do you intend on representing the final conceptual model? This will likely depend on the method chosen to elicit the conceptual model.

2.2 Explain critical conceptual design decisions

Preregistration Item

List and explain critical conceptual design decisions (Grimm et al., 2014), including:

- ☐ spatial and temporal scales,
- ☐ selection of entities and processes,
- ☐ representation of stochasticity and heterogeneity,
- ☐ consideration of local versus global interactions, environmental drivers, etc.
- ☐ Explain and justify the influence of particular theories, concepts, or earlier models against alternative conceptual design decisions that might lead to alternative model structures.

2.3 Model assumptions and uncertainties

Preregistration Item

Specify key assumptions and uncertainties underlying the model design, describing how uncertainty and variation will be represented in the model (Moallemi et al., 2019). Sources of uncertainty may include:

- ☐ exogenous uncertainties affecting the system,
- ☐ parametric uncertainty in input data and
- ☐ structural / conceptual nonparametric uncertainty in the model.

2.4 Identify predictor and response variables

Explanation

The identification and definition of primary model input variables should be driven by scenario definitions, and by the scope of the model described in the problem formulation phase (Mahmoud et al., 2009).

Preregistration Item

Identify and define system system variables and structures, referencing scenario definitions, and the scope of the model as described within problem formulation:

- ☐ What variables would support taking this action or making this decision?
- ☐ What additional variables may interact with this system (things we can't control, but can hopefully measure)?
- ☐ What variables have not been measured, but may interact with the system (often occurs in field or observational studies)?
- ☐ What variables are index or surrogate measures of variables that we cannot or have not measured?
- ☐ In what ways do we expect these variables to interact (model structures)?
- ☐ Explain how any key concepts or terms within problem or decision-making contexts, such as regulatory terms, will be operationalised and defined in a biologically meaningful way to answer the research question appropriately?

2.5 Define prior knowledge, data specification and evaluation

Explanation

This section specifies the plan for collecting, processing and preparing data available for parameterisation, determining model structure, and for scenario analysis. It also allows the researchers to disclose any prior interaction with the data.

2.5.1 Collate available data sources that could be used to parameterise or structure the model

Preregistration Item

For pre-existing data (delete as appropriate):

- ☐ Document the identity, quantity and provenance of any data that will be used to develop, identify and test the model.
- ☐ For each dataset, is the data open or publicly available?
- ☐ How can the data be accessed? Provide a link or contact as appropriate, indicating any restrictions on the use of data.
- ☐ Date of download, access, or expected timing of future access.
- ☐ Describe the source of the data - what entity originally collected this data? (National Data Set, Private Organisational Data, Own Lab Collection, Other Lab Collection, External Contractor, Meta-Analysis, Expert Elicitation, Other).
- ☐ Codebook and meta-data. If a codebook or other meta-data is available, link to it here and / or upload the document(s).
- ☐ Prior work based on this dataset - Have you published / presented any previous work based on this dataset? Include any publications, conference presentations (papers, posters), or working papers (in-prep, unpublished, preprints) based on this dataset you have worked on.
- ☐ Unpublished Prior Research Activity - Describe any prior but unpublished research activity using these data. Be specific and transparent.
- ☐ Prior knowledge of the current dataset - Describe any prior knowledge of or interaction with the dataset before commencing this study. For example, have you read any reports or publications about this data?
- ☐ Describe how the data is arranged, in terms of replicates and covariates.

Sampling Plan (for data you will collect, delete as appropriate):

- ☐ Data collection procedures - describe your data collection process, including how sites and transects or any other physical unit were selected and arranged. Describe any inclusion or exclusion rules, and the study timeline.
- ☐ Sample Size - Describe the sample size of your study.
- ☐ Sample Size Rationale - Describe how you determined the appropriate sample size for your study. It could include feasibility constraints, such as time, money or personnel.
- ☐ If sample size cannot be specified, specify a stopping rule - i.e. how will you decide when to terminate your data collection?

2.5.2 Data Processing and Preparation

Preregistration Item

- ☐ Describe any data preparation and processing steps, including manipulation of environmental layers (e.g. standardisation and geographic projection) or variable construction (e.g. Principal Component Analysis).

2.5.3 Describe any data exploration or preliminary data analyses

Explanation

In many modelling cases, it is necessary to perform preliminary analyses to understand the data and check that assumptions and requirements of the chosen modelling procedures are met. Data exploration prior to model fitting or development may include exploratory analyses to check for collinearity, spatial and temporal coverage, quality and resolution, outliers, or the need for transformations (Yates et al., 2018).

Caution! Beware questionable research practices and overfitting

Any modelling and analysis decisions based on data exploration or preliminary investigations increase the risk of questionable research practices and overfitting (Gould, Fraser, et al., 2026). Such *data-dependent* analytic decisions (Y. Liu et al., 2020) should ideally be performed using a small, independent subset of the full dataset in order to avoid model overfitting arising from "data leakage" or "model selection bias" (Arnqvist, 2020; Campbell, 2021; Kapoor & Narayanan, 2023).
The preregistration item below follows a protocol for registering 'flexible' analyses, please see our 'Primer on Adaptive Preregistration' in Gould et al. (2026) for details.

Preregistration Item

For each separate preliminary or investigatory analysis:

- ☐ State what quantity or estimand needs to be known in order to proceed with further decision-making about the modelling procedure, stating why the analysis is necessary.
- ☐ Explain how you will implement this analysis, as well as any techniques you will use to summarise and explore your data.
- ☐ What method will you use to represent this analysis (graphical, tabular, or otherwise, describe)?
- ☐ Specify exactly which parts of the data will be used, ensuring that any data subsets used are described in Section 4.1.1.
- ☐ Describe how the results will be interpreted, listing each potential analytic decision, as well as the analysis finding that will trigger each decision, where possible.

2.5.4 Data evaluation, exclusion and missing data

Explanation

Documenting issues with reliability is important because data quality and ecological relevance might be constrained by measurement error, inappropriate experimental design, and heterogeneity and variability inherent in ecological systems (Grimm et al., 2014). Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand (Mahmoud et al., 2009).

Preregistration Item

- ☐ Describe how you will determine how reliable the data is for the given model purpose. Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand
- ☐ Document any issues with data reliability.
- ☐ How will you determine what data, if any, will be excluded from your analyses?
- ☐ How will outliers be handled? Describe rules for identifying outlier data, and for excluding a site, transect, quadrat, year or season, species, trait, etc.
- ☐ How will you identify and deal with incomplete or missing data?

2.6 Conceptual model evaluation

Preregistration Item

- ☐ Describe how your conceptual model will be critically evaluated. Evaluation includes both the completeness and suitability of the overall model structure.
- ☐ How will you critically assess any simplifying assumptions (Augusiak et al., 2014)?
- ☐ Will this process will include consultation or feedback from a client, manager, or model user.

3 Formalise and Specify Model

Explanation

In this section describe what quantitative methods you will use to build the model/s, explain how they are relevant to the client/manager/user's purpose.

3.1 Model class, modelling framework and approach

Explanation

Modelling approaches can be described as occurring on a spectrum from correlative or phenomenological to mechanistic or process-based (Yates et al., 2018); where correlative models use mathematical functions fitted to data to describe underlying processes, and mechanistic models explicitly represent processes and details of component parts of a biological system that are expected to give rise to the data (White & Marshall, 2019). A model 'class,' 'family' ' or 'type' is often used to describe a set of models each of which has a distinct but related sampling distribution (C. C. Liu & Aitkin, 2008). The model family is driven by choices about the types of variables covered and the nature of their treatment, as well as structural features of the model, such as link functions, spatial and temporal scales of processes and their interactions (Jakeman et al., 2006).

Preregistration Item

- ☐ Describe what modelling framework, approach or class of model you will use to implement your model and relate your choice to the model purpose and analytical objectives described in Section 1.1.2 and Section 1.1.3.

3.2 Choose model features and family

Explanation

All modelling approaches require the selection of model features, which conform with the conceptual model and data specified in previous steps (Jakeman et al., 2006). The choice of model are determined in conjunction with features are selected. Model features include elements such as the functional form of interactions, data structures, measures used to specify links, any bins or discretisation of continuous variables. It is usually difficult to change fundamental features of a model beyond an early stage of model development, so careful thought and planning here is useful to the modeller (Jakeman et al., 2006). However, if changes to these fundamental aspects of the model do need to change, document how and why these choices were made, including any results used to support any changes in the model.

3.2.1 Operationalising Model Variables

Preregistration Item

- ☐ For each response, predictor, and covariate, specify how these variables will be operationalised in the model. This should relate directly to the analytical and/or management objectives specified during the problem formulation phase. Operationalisations could include: the extent of a response, an extreme value, a trend, a long-term mean, a probability distribution, a spatial pattern, a time-series, qualitative change, such as a direction of change or, the frequency, location, or probability of some event occurring. Specify any treatment of model variables, including whether they are lumped / distributed, linear / non-linear, stochastic / deterministic (Jakeman et al., 2006).
- ☐ Provide a rationale for your choices, including why plausible alternatives under consideration were not chosen, and relate your justification back to the purpose, objectives, prior knowledge and or logistical constraints specified in the problem formulation phase (Jakeman et al., 2006).

3.2.2 Choose model family

Preregistration Item

- ☐ Specify which family of statistical distributions you will use in your model, and describe any transformations, or link functions.
- ☐ Include in your rational for selection, detail about which variables the model outputs are likely sensitive to, what aspects of their behaviour are important, and any associated spatial or temporal dimensions in sampling.

3.3 Describe *approach* for identifying model structure

Explanation

This section relates to the process of determining the best/most efficient/parsimonious representation of the system at the appropriate scale of concern (Jakeman et al., 2006) that best meets the analytical objectives specified in the problem formulation phase. Model structure refers to the choice of variables included in the model, and the nature of the relationship among those variables. Approaches to finding model structure and parameters may be knowledge-supported, or data-driven (Boets et al., 2015). Model selection methods can include traditional inferential approaches such as unconstrained searches of a dataset for patterns that explain variations in the response variable, or use of ensemble-modelling methods (Barnard et al., 2019). Ensemble modelling procedures might aim to derive a single model, or a multi-model average (Yates et al., 2018). Refining actions to develop a model could include iteratively dropping parameters or adding them, or aggregating / disaggregating system descriptors, such as dimensionality and processes (Jakeman et al., 2006).

Preregistration Item

- ☐ Specify what approach and methods you will use to identify model structure and parameters.
- ☐ If using a knowledge-supported approach to deriving model structure (either in whole or in part), specify model structural features, including:
  - the functional form of interactions (if any)
  - data structures,
  - measures used to specify links,
  - any bins or discretisation of continuous variables (Jakeman et al., 2006),
  - any other relevant features of the model structure.

3.4 Describe parameter estimation technique and performance criteria

Explanation

Before calibrating the model to the data, the performance criteria for judging the calibration (or model fit) are specified. These criteria and their underlying assumptions should reflect the desired properties of the parameter estimates / structure (Jakeman et al., 2006). For example, modellers might seek parameter estimates that are robust to outliers, unbiased, and yield appropriate predictive performance. Modellers will need to consider whether the assumptions of the estimation technique yielding those desired properties are suited to the problem at hand. For integrated or sub-divided models, other considerations might include choices about where to disaggregate the model for parameter estimation; e.g. spatial sectioning (streams into reaches) and temporal sectioning (piece-wise linear models) (Jakeman et al., 2006).

3.4.1 Parameter estimation technique

Preregistration Item

- ☐ Specify what technique you will use to estimate parameter values, and how you will supply non-parametric variables and/or data (e.g. distributed boundary conditions). For example, will you calibrate all variables simultaneously by optimising fit of model outputs to observations, or will you parameterise the model in a piecemeal fashion by either direct measurement, inference from secondary data, or some combination (Jakeman et al., 2006).
- ☐ Identify which variables will be parameterised directly, such as by expert elicitation or prior knowledge.
- ☐ Specify which algorithm(s) you will use for any data-driven parameter estimation, including supervised, or unsupervised machine learning, decision-tree, K-nearest neighbour or cluster algorithms (Z. Liu et al., 2018).

3.4.2 Parameter estimation / model fit performance criteria

Preregistration Item

- ☐ Specify which suite of performance criteria you will use to judge the performance of the model. Examples include correlation scores, coefficient of determination, specificity, sensitivity, AUC, etcetera (Yates et al., 2018).
- ☐ Relate any underlying assumptions of each criterion to the desired properties of the model, and justify the choice of performance metric in relation
- ☐ Explain how you will identify which model features or components are significant or meaningful.

3.5 Model assumptions and uncertainties

Preregistration Item

- ☐ Specify assumptions and key uncertainties in the formal model. Describe what gaps exist between the model conception, and the real-world problem, what biases might this introduce and how might this impact any interpretation of the model outputs, and what implications are there for evaluating model-output to inform inferences or decisions?

3.6 Specify formal model(s)

Explanation

Once critical decisions have been made about the modelling approach and method of model specification, the conceptual model is translated into the quantitative model.

Preregistration Item

- ☐ Specify all formal models
  - ☐ Note, For data-driven approaches to determining model structure and or parameterisation, it may not be possible to respond to this preregistration item. In such cases, explain why this is the case, and how you will document the model(s) used in the final analysis.
- ☐ For quantitative model selection approaches, including ensemble modelling, specify each model used in the candidate set, including any null or full/global model.

4 Model Calibration, Validation & Checking

4.1 Model calibration and validation scheme

Explanation

This section pertains to any data calibration, validation or testing schemes that will be implemented. For example, the model may be tested on data independent of those used to parameterise the model (external validation), or the model may be cross-validated on random sub-samples of the data used to parameterise the model (internal cross-validation Yates et al., 2018; Barnard et al., 2019). For some types of models, hyper-parameters are estimated from data, and may be tuned on further independent holdouts of the training data ("validation data").

Preregistration Item

- ☐ Describe any data calibration, validation and testing scheme you will implement, including any procedures for tuning or estimating model hyper-parameters (if any).

4.1.1 Describe calibration/validation data

Explanation & Rationale

The following items pertain to properties of the *datasets* used for calibration (training), validation, and testing.

Preregistration Item

If partitioning data for cross-validation or similar approach (delete as needed):

- ☐ Describe the approach specifying the number of folds that will be created, the relative size of each fold, and any stratification methods used for ensuring evenness of groups between folds and between calibration / validation data?

If using external / independent holdout data for model testing and evaluation (delete as needed):

- ☐ Which data will be used as for the testing data? What method will be used for generating training / test data subsets?
- ☐ Describe any known differences between the training/validation and testing datasets, the relative size of each, as well as any stratification methods used for ensuring evenness of groups between data sets?
- ☐ It is preferable that any independent data used for model testing remains unknown to modellers during the process of model development, please describe the relationship modellers have to model validation data, will independent datasets be known or accessible to any modeller or analyst?

If performing any data exploration or other preliminary analyses (delete as needed)

- ☐ Describe any data subsets used for each analysis in *Section 2.5.3*.

4.2 Implementation verification

Explanation & Examples

Model implementation verification is the process of ensuring that the model has been correctly implemented, and that the model performs as described by the model description (Grimm et al., 2014). This process is distinct from model checking, which assesses the model's performance in representing the system of interest (Conn et al., 2018).

- Checks for verification implementation should include i) thoroughly checking for bugs or programming errors, and ii) whether the implemented model performs as described by the model description (Grimm et al., 2014).
- Qualitative tests could include syntax checking of code, and peer-code review (Ivimey et al., 2023). Technical measures include using unit tests, or in-built checks within functions to prevent potential errors.

Preregistration Item

☐ What Quality Assurance measures will you take to verify the model has been correctly implemented? Specifying a priori quality assurance tests for implementation verification may help to avoid selective debugging and silent errors.

4.3 Model checking

Rationale & Explanation

"Model Checking" goes by many names ("conditional verification", "quantitative verification", "model output verification"), and refers to a series of analyses that assess a model's performance in representing the system of interest (Conn et al., 2018). Model checking aids in diagnosing assumption violations, and reveals where a model might need to be altered to better represent the data, and therefore system (Conn et al., 2018). Quantitative model checking diagnostics include goodness of fit, tests on residuals or errors, such as for heteroscedascity, cross-correlation, and autocorrelation (Jakeman et al., 2006).

4.3.1 Quantitative model checking

Preregistration Item

During this process, observed data, or data and patterns that guided model design and calibration, are compared to model output in order to identify if and where there are any systematic differences.

☐ Specify any diagnostics or tests you will use during model checking to assess a model's performance in representing the system of interest.

☐ For each test, specify the criteria that will you use to interpret the outcome of the test in assessing the model's ability to sufficiently represent the gathered data used to develop and parameterise the model.

4.3.2 Qualitative model checking

Explanation

This step is largely informal and case-specific, but requires, 'face validation' with model users / clients / managers who aren't involved in the development of the model to assess whether the interactions and outcomes of the model are feasible an defensible (Grimm et al., 2014). This process is sometimes called a "laugh test" or a "pub test" and in addition to checking the model's believability, it builds the client's confidence in the model (Jakeman et al., 2006). Face validation could include structured walk-throughs, or presenting descriptions, visualisations or summaries of model results to experts for assessment.

Preregistration Item

☐ Briefly explain how you will qualitatively check the model, and whether and how you will include users and clients in the process.

4.3.3 Assumption Violation Checks

Preregistration Item

The consequences of assumption violations on the interpretation of results should be assessed (Araújo et al., 2019).

☐ Explain how you will demonstrate robustness to model assumptions and check for violations of model assumptions.
☐ If you cannot perform quantitative assumption checks, describe what theoretical justifications would justify a lack of violation of or robustness to model assumptions.
☐ If you cannot demonstrate or theoretically justify violation or robustness to assumptions, explain why not, and specify whether you will discuss assumption violations and their consequences for interpretation of model outputs.
☐ If assumption violations cannot be avoided, explain how you will explore the consequences of assumption violations on the interpretation of results (To be completed in interim iterations of the preregistration, only if there are departures from assumptions as demonstrated in the planned tests above).

5 Model Validation and Evaluation

Explanation

The model validation & evaluation phase comprises a suite of analyses that collectively inform inferences about whether, and under what conditions, a model is suitable to meet its intended purpose (Augusiak et al., 2014). Errors in design and implementation of the model and their implication on the model output are assessed. Ideally independent data is used against the model outputs to assess whether the model output behaviour exhibits the required accuracy for the model's intended purpose. The outcomes of these analyses build confidence in the model applications and increase understanding of model strengths and limitations. Model evaluation including, model analysis, should complement model checking. It should evaluate model checking, and consider over-fitting and extrapolation. As the proportion of calibrated or uncertain parameters increases, so does the risk that the model seemingly works correctly, but for the wrong mechanistic reasons (Boettiger, 2022). Evaluation thus complements model checking because we can rule out the chance that the model fits the calibration data well, but has not captured the relevant ecological mechanisms of the system pertinent to the research question or the decision problem underpinning the model (Grimm et al., 2014). Evaluation of model outputs against external data in conjunction with the results from model checking provide information about the structural realism and therefore credibility of the model (Grimm & Berger, 2016).

5.1 Model output corroboration

Explanation

Ideally, model outputs or predictions are compared to independent data and patterns that were not used to develop, parameterise, or verify the model. Testing against a dataset of response and predictor variables that are spatially and/or temporally independent from the training dataset minimises the risk of artificially inflating model performance measures (Araújo et al., 2019). Although the corroboration of model outputs against an independent validation dataset is considered the 'gold standard' for showing that a model properly represents the internal organisation of the system, model validation is not always possible because empirical experiments are infeasible or model users are working on rapid-response time-frames, hence, why ecologists often model in the first place (Grimm et al., 2014). Independent predictions might instead be

tested on sub-models. Alternatively, patterns in model output that are robust and seem characteristic of the system can be identified and evaluated in consultation with the literature or by experts to judge how accurate the model output is (Grimm et al., 2014).

Preregistration Item

- ☐ State whether you will corroborate the model outputs on external data, and document any independent validation data in step.
- ☐ It is preferable that any independent data used for model evaluation remains unobserved to modellers during the process of model building (Dwork et al., 2015), describe the relationship modellers have to model validation data, e.g. will independent datasets be known to any modeller or analyst involved in the model building process?
- ☐ If unable to evaluate the model outputs against independent data, explain why and explain what steps you will take to interrogate the model.

5.2 Choose performance metrics and criteria

Explanation

Model performance can be quantified by a range of tests, including measures of agreement between predictions and independent observations, or estimates of accuracy, bias, calibration, discrimination refinement, resolution and skill (Araújo et al., 2019). Note that the performance metrics and criteria in this section are used for evaluating the structured and parameterised models (ideally) on independent holdout data, so this step is additional to any performance criteria used for determining model structure or parameterisation (Section 3.4.2).

Preregistration Item

- ☐ Specify what performance measures you will use to evaluate the model and briefly explain how each test relates to different desired properties of a model's performance.
- ☐ Spatial, temporal and environmental pattern of errors and variance can change the interpretation of model predictions and conservation decisions (Araújo et al., 2019), where relevant and possible, describe how you will characterise and report the spatial, temporal and environmental pattern of errors and variance.
- ☐ If comparing alternative models, specify what measures of model comparison or out-of-sample performance metrics will you use to find support for alternative models or else to optimise predictive ability. State what numerical threshold or qualities you will use for each of these metrics.

5.3 Model analysis

Rationale & Explanation

Uncertainty in models arises due to incomplete system understanding (which processes to include, or which interact), from imprecise, finite and sparse data measurements, and from uncertainty in input conditions and scenarios for model simulations or runs (Jakeman et al., 2006). Non-technical uncertainties can also be introduced throughout the modelling process, such as uncertainties arising from issues in problem-framing, indeterminicies, and modeller / client values (Jakeman et al., 2006).
The purpose of model analysis is to prevent blind trust in the model by understanding how model outputs have emerged, and to 'challenge' the model by verifying whether the model is still believable and fit for purpose if one or more parameters are changed (Grimm et al., 2014).
Model analysis should increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system. Model analysis typically consists of sensitivity analyses preceded by uncertainty analyses (Saltelli et al., 2019), and a suite of other simulation or other computational experiments. The aim of such computational experiments is to increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system (Grimm et al., 2014). Uncertainty analyses and sensitivity

analyses augment one another to draw conclusions about model uncertainty.
Because the results from a full suite of sensitivity analysis and uncertainty analysis can be difficult to interpret due to the number and complexity of causal relations examined (Jakeman et al., 2006), it is useful for the analyst to relate the choice of analysis to the modelling context, purpose and analytical objectives defined in the problem formulation phase, in tandem with any critical uncertainties that have emerged during model development and testing prior to this point.

5.3.1 Uncertainty Analyses

Explanation

Uncertainty can arise from different modelling techniques, response data and predictor variables (Araújo et al., 2019). Uncertainty analyses characterise the uncertainty in model outputs, and identify how uncertainty in model parameters affects uncertainty in model output, but does not identify which model assumptions are driving this behaviour (Grimm et al., 2014; Saltelli et al., 2019). Uncertainty analyses can include propagating known uncertainties through the model, or by investigating the effect of different model scenarios with different parameters and modelling technique combinations (Araújo et al., 2019), for example. It could also include characterising the output distribution, such as through empirical construction using model output data points. It could also include extracting summary statistics like the mean, median and variance from this distribution, and perhaps constructing confidence intervals on the mean (Saltelli et al., 2019).

Preregistration Item

- ☐ Describe how you will characterise model and data uncertainties, e.g. propagating known uncertainties through the model, investigating the effect of different model scenarios with different parameters and modelling technique combinations (Araújo et al., 2019), or empirically constructing model distributions from model output data points, and extracting summary statistics, including the mean, median, variance, and constructing confidence intervals (Saltelli et al., 2019).
- ☐ Relate your choice of analysis to the context and purposes of the model described in the problem formulation phase. For instance , discrepancies between model output and observed output may be important for forecasting models, where cost, benefit, an risk over a substantial period must be gauged, but much less critical for decision-making or management models where the user may be satisfied with knowing that the predicted ranking order of impacts of alternative scenarios or management options is likely to be correct, with only a rough indication of their sizes" (Jakeman et al., 2006).
- ☐ Briefly describe how you will summarise the results of these in silico experiments with graphical, tabular, or other devices, such as summary statistics.
- ☐ If the chosen modelling approach is able to explicitly articulate uncertainty due to data, measurements or baseline conditions, such as by providing estimates of uncertainty (typically in the form of probabilistic parameter covariance, Jakeman et al., 2006), specify which measure of uncertainty you will use.

5.3.2 Sensitivity analyses

Explanation

Sensitivity analysis examines how uncertainty in model outputs can be apportioned to different sources of uncertainty in model input (Saltelli et al., 2019).

Preregistration Item

- ☐ Describe the sensitivity analysis approach you will take: deterministic sensitivity, stochastic sensitivity (variability in the model), or scenario sensitivity (effect of changes based on scenarios).
- ☐ Describe any sensitivity analyses you will conduct by specifying which parameters will be held constant, which will be varied, and the range and intervals of values over which those parameters

will be varied.
- ☐ State the primary objective of each sensitivity analysis, for example, to identify which input variables contribute the most to model uncertainty so that these variables can be targeted for further data collection, or alternatively to identify which variables or factors contribute little to overall model outputs, and so can be 'dropped' from future iterations of the model (Saltelli et al., 2019).

5.3.3 Model application or scenario analysis

Preregistration Item

- ☐ Specify any input conditions and relevant parameter values for initial environmental conditions and decision-variables under each scenario specified in Section 1.
- ☐ Describe any other relevant technical details of model application, such as methods for how you will implement any simulations or model projections.
- ☐ What raw and transformed model outputs will you extract from the model simulations or projections, and how will you map, plot, or otherwise display and synthesise the results of scenario and model analyses.
- ☐ Explain how you will analyse the outputs to answer your analytical objectives. For instance, describe any trade-off or robustness analyses you will undertake to help evaluate and choose between different alternatives in consultation with experts or decision-makers.

5.3.4 Other simulation experiments / robustness analyses

Preregistration Item

- ☐ Describe any other simulation experiments, robustness analyses or other analyses you will perform on the model, including any metrics and their criteria / thresholds for interpreting the results of the analysis.

References

Araújo, M., Anderson, R., Márcia Barbosa, A., Beale, C., Dormann, C., Early, R., Garcia, R., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R., Zimmermann, N., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Sci Adv*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Arnqvist, G. (2020). Mixed models offer no freedom from degrees of freedom. *Trends in Ecology & Evolution*, *35*(4), 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Barnard, D. M., Germino, M. J., Pilliod, D. S., Arkle, R. S., Applestein, C., Davidson, B. E., & Fisk, M. R. (2019). Can't see the random forest for the decision trees: Selecting predictive models for restoration ecology. *Restoration Ecology*. https://doi.org/10.1111/rec.12938

Boets, P., Landuyt, D., Everaert, G., Broekx, S., & Goethals, P. L. M. (2015). *Evaluation and comparison of data-driven and knowledge-supported bayesian belief networks to assess the habitat suitability for alien macroinvertebrates*. *74*, 92–103. https://doi.org/10.1016/j.envsoft.2015.09.005

Boettiger, C. (2022). The forecast trap. *Ecol Lett*. https://doi.org/10.1111/ele.14024

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210x.13559

Cartwright, S. J., Bowgen, K. M., Collop, C., Hyder, K., Nabe-Nielsen, J., Stafford, R., Stillman, R. A., Thorpe, R. B., & Sibly, R. M. (2016). Communicating complex ecological models to non-scientist end users. *Ecological Modelling*, *338*, 51–59. https://doi.org/10.1016/j.ecolmodel.2016.07.012

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *9*, 341–317. https://doi.org/10.1002/ecm.1314

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Fraser, H., Rumpff, L., Yen, J. D. L., Robinson, D., & Wintle, B. A. (2017). Integrated models to support multiobjective ecological restoration decisions. *Conservation Biology*, *31*(6), 1418–1427. https://doi.org/10.

1111/cobi.12939

Gould, E., Fraser, H., Wintle, B., Rumpff, L., & Fidler, F. (2026). *A framework for questionable research practices in ecological modelling*. https://doi.org/10.32942/X2DQ0K

Gould, E., Jones, C., Yen, J. D. L., Fraser, H., Wootton, H., Good, M., Duncan, D., Rumpff, L., & Fidler, F. (2026). *EcoConsPreReg: A guide to adaptive preregistration for model-based research in ecology and conservation*. https://doi.org/10.5281/ZENODO.10807029

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129–139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., & Berger, U. (2016). Structural realism, emergence, and predictions in next-generation ecological modelling: Synthesis from a special issue. *Ecological Modelling*, *326*, 177–187. https://doi.org/10.1016/j.ecolmodel.2016.01.001

Ivimey, E. R., Pick, J. L., Bairos, K. R., Culina, A., Gould, E., Grainger, M., Marshall, B. M., Moreau, D., Paquet, M., Royauté, R., Sánchez, A., Silva, I., & Windecker, S. M. (2023). Implementing code review in the scientific workflow: Insights from ecology and evolutionary biology. *Journal of Evolutionary Biology*. https://doi.org/10.1111/jeb.14230

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Liu, Y., Althoff, T., & Heer, J. (2020). *Proceedings of the 2020 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3313831.3376533

Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., & Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: Recent progress and future challenges. *Environmental Reviews*, *26*(4), 339–350. https://doi.org/10.1139/er-2018-0034

Mahmoud, M., Liu, Y., Hartmann, H., Stewart, S., Wagener, T., Semmens, D., Stewart, R., Gupta, H., Dominguez, D., Dominguez, F., Hulse, D., Letcher, R., Rashleigh, B., Smith, C., Street, R., Ticehurst, J., Twery, M., Delden, H. van, Waldick, R., … Winter, L. (2009). A formal framework for scenario development in support of environmental decision-making. *Environmental Modelling & Software*, *24*(7), 798–808. https://doi.org/10.1016/j.envsoft.2008.11.010

McCarthy, M. A., Thompson, C. J., Moore, A. L., & Possingham, H. P. (2011). Designing nature reserves in the face of uncertainty. *Ecology Letters*, *14*(5), 470–475. https://doi.org/10.1111/j.1461-0248.2011.01608.x

McDonald-Madden, E., Baxter, P. W. J., & Possingham, H. P. (2008). Making robust decisions for conservation with restricted money and knowledge. *Journal of Applied Ecology*, *45*(6), 1630–1638. https://doi.org/10.1111/j.1365-2664.2008.01553.x

Moallemi, E. A., Elsawah, S., & Ryan, M. J. (2019). Strengthening "good" modelling practices in robust decision support: A reporting guideline for combining multiple model-based methods. *Mathematics and Computers in Simulation*. https://doi.org/10.1016/j.matcom.2019.05.002

Moon, K., Guerrero, A. M., Adams, Vanessa. M., Biggs, D., Blackman, D. A., Craven, L., Dickinson, H., & Ross, H. (2019). Mental models for conservation research and practice. *Conservation Letters*, *12*(3), e12642. https://doi.org/10.1111/conl.12642

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29–39. https://doi.org/10.1016/j.envsoft.2019.01.012

White, C. R., & Marshall, D. J. (2019). Should we care if models are phenomenological or mechanistic. *Trends in Ecology & Evolution*, *34*(4), 276–278. https://doi.org/10.1016/j.tree.2019.01.006

Yates, K., Bouchet, P., Caley, M., Mengersen, K., Randin, C., Parnell, S., Fielding, A., Bamford, A., Ban, S., Barbosa, A., Dormann, C., Elith, J., Embling, C., Ervin, G., Fisher, R., Gould, S., Graf, R., Gregr, E., Halpin, P., … Sequeira, A. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol. (Amst.)*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

D.5 Case Study Preregistration

Below the case study preregistration is reproduced, which can also be accessed from github.com/egouldo/VEFMAP_VEG_Stage6 (see README for details).

Note that the version of the template used in the case study differs from the template in Section D.4. Following reflection on the case study, we refined the template further.

Preregistration for Vegetation Responses to Environmental Flows

Chris Jones[2], Elliot Gould[1], Jian Yen[2], Henry Wootton[2]

[1]School of Agriculture, Food and Ecosystem Sciences, University of Melbourne
[2]Arthur Rylah Institute for Environmental Research, Department of Energy, Environment and Climate Action

2024-03-25

Table of contents

1 Problem Formulation **1**
1.1 Model Context and Purpose . . . 2

2 Define Conceptual Model **5**
2.1 Choose elicitation and representation method . . . 5
2.2 Explain Critical Conceptual Design Decisions . . . 6
2.3 Model assumptions and uncertainties . . . 6
2.4 Identify predictor and response variables . . . 8
2.5 Define prior knowledge, data specification and evaluation . . . 9
2.6 Conceptual model evaluation . . . 15

3 Formalise and Specify Model **15**
3.1 Model class, modelling framework and approach . . . 15
3.2 Choose model features and family . . . 16
3.3 Describe *approach* for identifying model structure . . . 17
3.4 Describe parameter estimation technique and performance criteria . . . 18
3.5 Model assumptions and uncertainties . . . 19
3.6 Specify formal model(s) . . . 20

4 Model Calibration, Validation & Checking **23**
4.1 Model calibration and validation scheme . . . 23
4.2 Implementation verification . . . 24
4.3 Model checking . . . 24

5 Model Validation and Evaluation **26**
5.1 Model output corroboration . . . 26
5.2 Choose performance metrics and criteria . . . 27
5.3 Model analysis . . . 27

References **29**

1 Problem Formulation

ℹ Rationale & Explanation

This section specifies the decision-making context in which the model will be used or the intended scope and context of conclusions. Important components include the decision maker and stakeholders (including experts) and their view on: i) the nature of the problem or decision addressed and how the scope of the modelling tool fits within the (broader) context (i.e. model purpose; ii) the spatial and temporal scales relevant to the decision context; iii) specified desired outputs; iv) role and inclusion in model development and testing; v) whether they foresee unacceptable outcomes that need to be represented in the model (i.e. as constraints), and; vi) what future scenarios does the model need to account for (noting this may be revised later). It should also provide a summary of the domain of applicability of the model, and

reasonable extrapolation limits (Grimm et al., 2014).

1.1 Model Context and Purpose

Rationale & Explanation

Defining the purpose of the model is critical because the model purpose influences choices at later stages of model development (Jakeman et al., 2006). Common model purposes in ecology include: gaining a better qualitative understanding of the target system, synthesising and reviewing knowledge, and providing guidance for management and decision-making (Jakeman et al., 2006). Note that modelling objectives are distinct from the analytical objectives of the model.
The scope of the model includes temporal and spatial resolutions, which should also be defined here (Mahmoud et al., 2009). Any external limitations on model development, analysis and flexibility should also be outlined in this section (Jakeman et al., 2006).

1.1.1 Key stakeholders and model users

Preregistration Item

Identify relevant interest groups:
- ☑ Who is the model for?
- ☑ Who is involved in formulating the model?
- ☑ How will key stakeholders be involved in model development?
- ☑ Describe the decision-making context in which the model will be used (if relevant).

This preregistration document relates to the data analysis of a study on vegetation responses to environmental flows. The study forms one component of a larger set of studies within Stage 6 of the Victorian Environmental Flows Monitoring and Assessment Program (VEFMAP) managed by DELWP and delivered through the Arthur Rylah Institute (ARI). The study is funded by the state of Victoria, with a hierarchy of clients ranging from the Premier, to the Water Minister, Dep. Secretary, and program managers within the DELWP Water and Catchments division. The project is also co-funded through the Murray Darling Basin Plan, with corresponding requirements to contribute findings to Basin Plan reporting outputs. Other key stakeholders for this work are the Victorian Environmental Water Holder (VEWH) who manage the environmental water entitlements throughout the state, and the Victorian Catchment Management Authorities (CMAs) that manage the delivery of environmental water along individual waterways. Findings from this research can then be used to directly influence: management decisions through the CMAs; water allocation through the VEWH; and water investment through DELWP, the State Government and the Murray Darling Basin Authority. Other stakeholders, including the public, water authorities and institutions such as Parks Victoria (among many others) have a minor active role within the study.

The model formulation is developed by the researchers at ARI, with guidance from the program managers at DELWP, external researchers and an Independent Review Panel of external researchers for VEFMAP. The general model structure was proposed by the ARI researchers, which was initially interrogated, modified and validated with the program manager and external input. A more formal development of the model formulation was conducted in a collaboration between ARI researchers and external researchers with specific statistical expertise.

The model outputs, including recommendations, will be shared with decision makers (funders and managers) to guide their respective decisions. Decisions will require multiple inputs, such as the outcomes of this study and others, regulatory frameworks, funding availability, physical and practical constraints, and others. This study does not contribute to those decisions outside of providing the study outputs.

The three primary users of the model outputs are: 1) funders of environmental water (i.e. state and federal governments) to understand the magnitude of the outcomes of their investment, and the limitations or barriers to benefits; 2) managers of environmental water (i.e. CMAs) to guide and improve management decisions for environmental benefits; and 3) other researchers working in the field of waterway flow management and environmental flows locally and internationally.

1.1.2 Model purpose, context and problem context

Preregistration Item

Briefly outline:

- ☑ the ecological problem,
- ☑ the decision problem (if relevant), including the decision-trigger and any regulatory frameworks relevant to the problem,
- ☑ how the model will address the problem, being clear about the scope of the model i.e. is the model addressing the whole problem, or part of it? Are there any linked problems that your model should consider?
- ☑ Ensure that you specify any focal taxa and study objectives.

The ecological problem is related to river regulation. Natural flow regimes are major drivers of ecological processes within waterways. Natural flow regimes are altered by water storage and extraction, i.e. regulation. Environmental flows aim to provide flow components in a waterway to replace natural components that have been removed from the flow regime, specifically to provide ecological benefit to multiple taxa and physical processes. The decision problem relates to maximising the effectiveness of the environmental water deliveries to achieve their objectives. Decision triggers can occur for the funding of environmental water, i.e. whether the outcomes are worth the investment or if the investment needs to be altered. Decision triggers can also occur for the management of environmental water, i.e. whether the outcomes suggest changes to the current management process. For this study, the evidence is provided for others to address those decisions. There are various regulatory frameworks that are relevant to the funding and delivery of environmental water, but they are not within the scope of this assessment as they are in the hands of the decision makers.

The model itself within this study aims to provide clear evidence of the influence of environmental flows, and other factors (particularly exotic vegetation and livestock grazing), on native vegetation (cover and diversity) within regulated river channels. Outputs will aim to give practical evidence and the implications of responses to directly improve and guide the management of environmental water delivery for native vegetation benefit. Given that the model will incorporate data from across a number of waterways across a large part of Victoria, the model applications will speak to each of those waterways, as well as potential extrapolation to waterways not surveyed. The models will not aim to provide explicit predictions within the waterways surveyed, or to specific un-surveyed waterways, but extension of these models for prediction will be highlighted for future investigation beyond this study.

1.1.3 Analytical objectives

Explanation

How will the model be analysed, what analytical questions will the model be used to answer? For example, you might be using your model in a scenario analysis to determine which management decision is associated with minimum regret or the highest likelihood of improvement. Other examples from ecological decision-making include: to compare the performance of alternative management actions under budget constraint (Fraser et al., 2017), to search for robust decisions under uncertainty (McDonald-Madden et al., 2008), to choose the conservation policy that minimises uncertainty (McCarthy et al., 2011). See other examples in Moallemi et al. (2019).

Preregistration Item

Provide detail on the analytical purpose and scope of the model:

- ☑ How will the model be analysed and what analytical questions will the model be used to answer?
- ☐ Candidate decisions should be investigated and are specified a priori. Depending on the modelling context, they may be specified by stakeholders, model users or the analyst (Moallemi et al., 2019).
  - ☐ Describe the method used to identify relevant management actions and
  - ☐ specify management actions to be considered included in the model.
  - ☐ Are there potentially unacceptable management or policy outcomes identified by stakeholders that should be captured in the model, i.e. as constraints?
- ☑ Are there scenarios that model inputs or outputs that must accommodated? Scenarios should be set a priori, (i.e. before the model is built, Moallemi et al., 2019) and may be stakeholder-defined or

driven by the judgement of the modeller or other experts (Mahmoud et al., 2009).
- ☐ If relevant, describe what processes you will use to elicit and identify relevant scenarios, e.g. literature review, structured workshops with stakeholders or decision-makers.
- ☐ Specify scenarios under which decisions are investigated.

The model objectives for this study are to provide evidence for vegetation responses to environmental flow delivery over short (<1yr) to medium (<5yr) term with consideration of additional significant factors, such as exotic vegetation cover and livestock grazing. These objectives can be captured in the following questions: what is the short-term vegetation response to a single environmental flow event? what is the vegetation response to repeated flow events that have been delivered in recent years? how much is the native vegetation response to environmental flows limited by the abundance of exotic vegetation or livestock grazing?

The models will aim to quantify vegetation responses (indicated by the level or change in plant cover or diversity) to environmental flows and to quantify the effect of exotic vegetation cover or livestock grazing presence on these responses. Vegetation cover data that will be the response variables to models (and exotic vegetation covariates) are collected on the individual species level but will likely be grouped into relevant response classes. These data also enable evaluation of vegetation diversity within relevant response classes. The specific groupings are yet to be determined. More information on these groupings is provided below in Section 2.4.

The models themselves will be analysed by assessing the model fit and parameters indicating suitability of model structure given the data. The specific tests to be used will depend on the model structure used. More information is provided in Section 3.2 and Section 2.3. Future study will investigate the predictive capacity of the models within and between different waterways to enable transferability of the data, but this is beyond the scope of the current piece of work and is not described further here.

Future research will involve testing scenarios of different flow regimes to what we have observed. This includes estimating the expected data we would have collected if the flows that were delivered were not delivered - thus effectively modelling a control or counterfactual dataset to the observed data. It also includes testing various future scenarios with different hypothetical future regimes in the medium (<5yr) or long (>5yr) term. However, again this is likely to be beyond the scope of the current study which will focus solely on the collected data.

1.1.4 Logistical Constraints

Preregistration Item

- ☑ What degree of flexibility is required from the model? Might the model need to be quickly reconfigured to explore new scenarios or problems proposed by clients / managers / model-users?
- ☑ Are there any limitations on model development analysis and flexibility, such as time or budget constraints, for example, does a model need to be deployed rapidly?
  - ☑ When must the model be completed by, e.g. to help make a decision?

The models will require a certain amount of flexibility to be able to change the time frames over which data are compared and the set of predictor variables that will change depending on the time frames selected. Funding constraints and client objectives may also require changes to the modelling process or purpose at any point. Currently there are no significant limitations on model development - sufficient time and funding resources - but this may change.

1.1.5 Model Scope, Scale and Resolution

Preregistration Item

- ☐ The choice of a model's boundaries is closely linked to the choice of how finely to aggregate the behaviour within the model (Jakeman et al., 2006)
  - what is the intended scale, and resolution of the model (temporal, spatial or otherwise)?
- ☐ Where is the boundary of the modelled system? Everything outside beyond the boundary and not crossing it is to be ignored within the domain of the model, and everything crossing the boundary is to be treated as external forcing (known/unknown), or else as model outputs (observed, or not, Jakeman et al., 2006).

Each of the models described in this preregistration is aggregated differently depending on the response variable and depending on the particular model specification.

Vegetation Cover Models

- Full Model: Models vegetation cover for any given transect
- Simplified Model 1: "flow regime model"
- Simplified Model 2: "flow events model"

Species Richness Models

1.1.6 Intended application of results

Explanation

Preregistration Items in this section are relevant to model transferability (Yates et al., 2018) and constraints on generality in model analysis interpretation. How far do can the results be extrapolated based on the study design (data + model + analysis)? For instance, if there are many confounding variables and not enough spatial / environmental replication, then making broader more general claims beyond the stated boundaries of the model (Section 1.1.3) may not be warranted. However, larger generalisations about results may be acceptable if the data comes from experimentally manipulated or controlled systems.

Preregistration Item

☑ What is the intended domain in which the model is to be applied? Are there any reasonable extrapolation limits beyond which you expect the model should not be applied (Grimm et al., 2014)?

The models developed in this study will be applied to only the river systems for which data was collected and used to fit these models. However, improved understanding about the ecological responses to environmental flows for the modelled systems may inform future conceptual models and statistical models based thereon.

Given that the primary goal of this study is to improve understanding about the effects of environmental flows on vegetation, these models resulting from this preregistration should not be directly used to inform predictions on which to base decisions about the management of environmental flows. The models developed in this study will instead be appropriately refined for use within a predictive modelling context.

2 Define Conceptual Model

Explanation

Conceptual models underpin the formal or quantitative model (Cartwright et al., 2016). The conceptual model describes the biological mechanisms relevant to the ecological problem and should capture basic premises about how the target system works, including any prior knowledge and assumptions about system processes. Conceptual models may be represented in a variety of formats, such as influence diagrams, linguistic model block diagram or bond graphs, and these illustrate how model drivers are linked to both outputs or observed responses, and internal (state) variables (Jakeman et al., 2006).

2.1 Choose elicitation and representation method

Preregistration Item

☑ Describe what method you will use to elicit or identify the conceptual model. Some common methods include interviews, drawings, and mapping techniques including influence diagrams, cognitive maps and Bayesian belief networks (Moon et al., 2019). It is difficult to decide and justify which method is most appropriate, see Moon et al. (2019) for guidance addressing this methodological question.
☑ Finally, how do you intend on representing the final conceptual model? This will likely depend on the method chosen to elicit the conceptual model.

We have developed a series of relevant conceptual models for this work over the past four years that build on previous published work by other researchers, as well as our own research, observations and many discussions.

These conceptual models are described in text within program reports and are summarised in diagrams. The models are biological/ecological only and do not include links to decision makers or values as these constraints are not within our capability to influence. Therefore, the models primarily summarise the dominant drivers of vegetation attributes (cover, diversity, distribution, composition) within waterways. Not all important drivers are included because it is impossible to account for everything within our study. Our final model for this particular study will be described in text and summarised within one or more non-quantitative diagrams.

2.2 Explain Critical Conceptual Design Decisions

Preregistration Item

List and explain critical conceptual design decisions (Grimm et al., 2014), including:

- ☑ spatial and temporal scales,
- ☑ selection of entities and processes,
- ☑ representation of stochasticity and heterogeneity,
- ☑ consideration of local versus global interactions, environmental drivers, etc.
- ☑ Explain and justify the influence of particular theories, concepts, or earlier models against alternative conceptual design decisions that might lead to alternative model structures.

This current study will focus on two different time frames: short term (months) in relation to before and after event surveys; as well as long-term (years) in relation to patterns resulting from previous years of particular flow regimes. The models therefore need to describe the short term responses of plants to events as well as the cumulative responses of plants to multiple types of events within years repeated over several years, i.e. regimes.

The entities will be the vegetation response variables (cover, diversity, composition and perhaps distribution) as well as various flow variables and other factors such as bank elevation, site, river system, rainfall and livestock grazing. The flow variables will depend on the time frame. Short term responses will include the presence and magnitude of a flow event and potentially the time of year and duration of the event. Long term responses will include the presence of individual flow events over the previous years (number of years TBC) and categorisation of the types of events that occurred, e.g. was a winter/spring flow natural or e-flow? Was it small or large? Categories TBC.

The primary process included is the inundation of plants by river flow. Rainfall to provide water will be included as a covariate rather than a process. Livestock grazing will also be evaluated as a damaging process through various mechanisms (plant consumption, trampling, soil compaction, nutrient addition, pugging, etc.) but these mechanisms are difficult to untangle so will most likely be combined if we can't do so adequately.

There is a large amount of heterogeneity in vegetation patterns within and between sites as well as stochasticity in the responses to events (and regimes). While both will play a large role in the modelling process of this study, neither is substantially addressed in the conceptual models for responses. One key component that has been considered though is the issue of habitat unsuitability for some plants in some areas, for example, there are many locations on a bank (such as very steep banks) where it is very difficult for some or any plants to occur, so these sites will have no or low cover or diversity regardless of flows. Because of this, we may consider models that evaluate change in cover only where plants occur at a sample in at least one survey - i.e. excluding sample locations with zeros through the whole dataset. This is defensible because locations where occupancy is impossible are not effective for evaluating flow responses, however, the frequency and distribution of these 'zero' samples would need to be described as well.

There are clear local and global interactions within our study. Local interactions are dominated by the species abundance and composition of plants within a sample area (plant interactions). Rainfall may also interact with flow events and would influence sites or groups of sites separately. Global interactions include the effect of season (time of year) on the responses, which is largely influenced by day length and temperatures.

2.3 Model assumptions and uncertainties

Preregistration Item

Specify key assumptions and uncertainties underlying the model design, describing how uncertainty and variation will be represented in the model (Moallemi et al., 2019). Sources of uncertainty may include:

- ☑ exogenous uncertainties affecting the system,
- ☑ parametric uncertainty in input data and

☑ structural / conceptual nonparametric uncertainty in the model.

Exogenous uncertainty

By 'exogenous uncertainties' I'm assuming this refers to uncertainties in any of the possible predictor variables for the model that are not direct treatment variables, even those that may not be used. There are a large number of potential exogenous uncertainties affecting the system because the surveys are done in the natural environment. These include:

- The spatial variability of rainfall. For rainfall to be accessible to a plant it needs to fall on or very near the plant in most cases. However, rainfall is spatially patchy and a rain event will affect different sites in an area differently. So our rain gauge data will only be an approximation of the actual rainfall at a specific site.
- Climate. Aside from rainfall, there are many climate elements that will influence plants, such as temperature, humidity and solar radiation (light). We can provide estimated values for some of these at sites but this will not account for the spatial and temporal variation in these factors that may influence our results.
- Soil properties. We do not have accurate (or any in some cases) information on soil type at each site. Nor do we have an expectation of how soils might influence plant attributes and responses to flows. We may be able to obtain basic information from spatial mapping of soil types, but this would be at a very coarse scale and would also not account for shallow surface soil deposits or variation through the soil profile. We have some data on soil types at some sites from soil cores but not all.
- Funghi. Funghi may be considered exogenous or endogenous, but it may have a role in determining plant patterns and responses to flows. We have no data on this, nor any specific expectations on how funghi may influence results.

Parametric uncertainty

I think this is referring to the other predictor variables in the model that are directly associated with treatments. These are primarily associated with flow but also grazing and potentially exotic plant cover.

- Flow elevation zones (if used). There is uncertainty in our GPS measurements of elevation of sub-transects and quadrats. There is also uncertainty in the flow elevation through time. Also, the flow elevation peak may only last for a brief period so the duration affected by the peak height may be very short. This means that the flow elevation zones that combine all of this information are uncertain. Even if we do not use zones, each of the input data listed here will be included and will be uncertain as indicated.
- Flow magnitude and duration. Flow magnitude (discharge) is recorded in most regulated streams at set gauges. The data from these gauges is usually calibrated by the data manager at some point after the raw data are entered. We have little idea of how much uncertainty there is in the uncorrected or corrected data that we obtain from online resources. Also, it is assumed that the magnitude (and timing and duration of flows) at the gauge is representative of the sites near that gauge, but we are unsure how much that may vary spatially.
- Flow velocity. Flow velocity can have a significant impact on plants if it is fast enough to damage or remove plants, or if it alters the soil. However the calculation of velocity at any given point on a river is difficult and varies at all points throughout the river cross section depending on the flow magnitude, flow height, channel form and obstacles. We do not have the capacity to calculate or even broadly approximate velocities at different flows so this variable cannot be included. However, we have seen very little field evidence of plant damage or removal from regulated flows on waterways and velocities during most regulated flows are typically not high on the bank where the plants are.
- Livestock grazing is specified as present or absent only at each site, not the intensity or duration of grazing. This adds a lot of uncertainty to this variable because the intensity (density of animals) and duration are important factors for the impact. However, this information is difficult to obtain from landholders or estimate from the site. We do have the animal ID though (cattle or sheep). Also, grazing impacts are spatially patchy depending on the ease of access - livestock don't impact steep slopes or inaccessible ledges or obstacles much compared to flatter or gentler slopes. So within a grazed site, the grazing impact may vary between transects.
- We will consider models for native plant cover that include exotic plant cover as a predictor. The categorisation of native and exotic is fairly well recognised in Victoria and should not pose an issue, but the categorisation of the plant group is variable (see Section 2.5.2). Also, the estimate of cover is uncertain.

Structural non-parametric uncertainty

Model uncertainty is a given. The models we will use will not fit the data perfectly but we will be able to quantify this uncertainty in our model estimates. Uncertainty will be captured for our vegetation input data based

on the number of samples we have within our hierarchical data structure. Uncertainty will not be accounted for in our uncertain variables such as rainfall or flow where we have no way of estimating this uncertainty. Unknown uncertainty will be somewhat captured in random effects but this will be an approximation only. More information about the model form and potential uncertainty is provided in Section 3 but is not contained within our conceptual model.

2.4 Identify predictor and response variables

Explanation

The identification and definition of primary model input variables should be driven by scenario definitions, and by the scope of the model described in the problem formulation phase (Mahmoud et al., 2009).

Preregistration Item

Identify and define system variables and structures, referencing scenario definitions, and the scope of the model as described within problem formulation:

- ☑ What variables would support taking this action or making this decision?
- ☑ What additional variables may interact with this system (things we can't control, but can hopefully measure)?
- ☑ What variables have not been measured, but may interact with the system (often occurs in field or observational studies)?
- ☑ What variables are index or surrogate measures of variables that we cannot or have not measured?
- ☑ In what ways do we expect these variables to interact (model structures)?
- ☑ Explain how any key concepts or terms within problem or decision-making contexts, such as regulatory terms, will be operationalised and defined in a biologically meaningful way to answer the research question appropriately?

Response Variables

The response variables will be associated with plant cover and diversity, which are the most commonly specified variables of interest within stated management objectives. These variables are also widely used within ecological studies and are directly transferable/translatable to many other studies. The specific variables used are likely to be:

- Plant cover (by species or species group) at a particular time
- Change in plant cover (by species or group) over a particular period of time relating to the occurrence of one or more flow events
- Plant diversity (in relevant species response groups) at a particular time
- Change in plant diversity (in relevant species response groups) over a particular period of time relating to the occurrence of one or more flow events

Predictor Variables

Based on our problem formulation, the expected predictor variables to use within our models are:

- Bank elevation (by zone or elevation measure GPS)
- Flow elevation, duration, time.
- Livestock grazing (likely presence or absence (but can provide approximate intensity category), potentially by grazer type sheep/cattle)
- Rainfall (potentially, but need to be careful about what period of rainfall data is relevant to the responses)
- Site variables (random effects for hierarchical structure: sub-transects/quadrats < transects < site < system/basin)
- Exotic vegetation cover or diversity (for models with native vegetation as the response)
- Time/season (the period over which the event has occurred, for regimes this will be standardised to a set number of years so will not be included as time but would be categorised into regime type TBC)

Each of these variables is important for influencing decisions about the use of environmental water to achieve benefits for (native) riparian vegetation. Additional variables that may be useful but are difficult to obtain at the appropriate spatial resolution (such as soils and funghi) are unlikely to be added. We expect there to be interactions between flow and bank elevation, hence the need to incorporate the elevation in our models. There may be interactions between bank elevation and livestock impacts due to the softer ground at the bank margin

being more susceptible to trampling. Exotic vegetation will also interact with bank elevation and flows. Rainfall in summer is likely to be more influential than rainfall in winter.

2.5 Define prior knowledge, data specification and evaluation

Explanation

This section specifies the plan for collecting, processing and preparing data available for parameterisation, determining model structure, and for scenario analysis. It also allows the researchers to disclose any prior interaction with the data.

2.5.1 Collate available data sources that could be used to parameterise or structure the model

Preregistration Item

For pre-existing data (delete as appropriate):

- ☑ Document the identity, quantity and provenance of any data that will be used to develop, identify and test the model.
- ☑ For each dataset, are the data open or publicly available?
- ☑ How can the data be accessed? Provide a link or contact as appropriate, indicating any restrictions on the use of data.
- ☐ Date of download, access, or expected timing offuture access.
- ☑ Describe the source of the data - what entity originally collected these data? (National Data Set, Private Organisational Data, Own Lab Collection, Other Lab Collection, External Contractor, Meta-Analysis, Expert Elicitation, Other).
- ☐ Codebook and meta-data. If a codebook or other meta-data are available, link to it here and / or upload the document(s).
- ☑ Prior work based on this dataset - Have you published / presented any previous work based on this dataset? Include any publications, conference presentations (papers, posters), or working papers (in-prep, unpublished, preprints) based on this dataset you have worked on.
- ☑ Unpublished Prior Research Activity - Describe any prior but unpublished research activity using these data. Be specific and transparent.
- ☑ Prior knowledge of the current dataset - Describe any prior knowledge of or interaction with the dataset before commencing this study. For example, have you read any reports or publications about these data?
- ☑ Describe how the data are arranged, in terms of replicates and covariates.

For this section, I have used the relevant points above from both of the options above given that we have collected the data ourselves.

- The data were collected over a four year period (2016-2020) as part of the Victorian Environmental Flows Monitoring and Assessment Program. The program outline and the methods are outlined in reports found at this website: https://www.ari.vic.gov.au/research/rivers-and-estuaries/assessing-benefits-of-water-for-the-environment
- Data were collected from 44 sites across Victoria with sampling year and number of surveys varying between sites, from 3 times in one 'water year' to 11 times in four 'water years'. Each site had 5-10 permanent transects established and each of those transects had a series of sub-transects at increasing bank elevations at which data were collected each survey. More detail on how the data were collected is available in the manual.
- The data are not yet publicly available but will be available from an online database at a later date as a requirement of publicly funded data. There will be standard conditions of use for the data respecting IP and research contributions.
- A metadata file has not yet been produced for all of the relevant data files, but the column names have been intuitively labelled. A file will be produced and stored with the data or in an accessible repository. *Task/issue to create metadata file*
- There have been many unpublished client reports produced based on small subsets of these data but with very limited data analysis. There have also been many oral presentations and various summary documents (flyers etc.) shared with relevant stakeholders. A summary including some preliminary analysis of these data is provided in this published client report, available at the above website (Tonkin et al., 2020). There are other published papers in development based on this program but not on this dataset.

- Existing knowledge of the data comes from collecting it ourselves, conducting thorough data checking and cleaning, producing output summaries (species lists, sample sizes, etc.) and conducting a preliminary analysis on a small subset of the data for the report highlighted above and a separate study on one time period from one river: Sutton, N., Houghton, J., Vietz, G., Jones, C., Mole, B., Morris, K., Gower, T. 2020. Influence of Intervalley Transfers (IVT) on the Riverbanks and Bank Vegetation of the Goulburn and Campaspe Rivers. Report by Streamology and Arthur Rylah Institute for the Department of Environment, Land, Water and Planning. June, 2020.

2.5.2 Data Processing and Preparation

Preregistration Item

☑ Describe any data preparation and processing steps, including manipulation of environmental layers (e.g. standardisation and geographic projection) or variable construction (e.g. Principal Component Analysis).

This analysis is relatively complex, due to the large dataset and the integration of different data sources (vegetation data, flow, transect elevation, site attributes). Each of these data sources requires some careful data processing and preparation for use within the models. Here we outline the major actions - listed within issue #23 in our repository.

GPS data correlation The specific relevance of this issue for flow elevation is detailed below, but here we summarise the general issues with GPS correlation. High-accuracy GPS points are required for three parts of this study: 1) vegetation sampling locations, 2) flow elevation (data loggers), 3) bank profiles and soil moisture loggers. In this study, there were various sources of these GPS points, with each source not corresponding exactly to the others. A calibration is then used to align all points at a site for each source. The data sources include:

- Points for flow loggers on Campaspe: Unimelb
- Points for veg survey locations on Campaspe: ARI using UPG equipment
- Points for all other veg survey locations: ARI using ARI equipment (equiv to UPG)
- Zero grade heights for permanent water loggers: Ventia

In all cases, we will calibrate to the levels of the vegetation survey locations, so that the flow data and vegetation data align.

Flow data correlation with site elevation Flow data in this study was processed externally by a hydrology consultant with data and site information provided by the ARI team. For each site surveyed, flow data were compiled and an output of flow level (elevation in AHD) was produced at regular intervals (interval time depending on data inputs). While all vegetation survey sites were attempted to have matching flow data, some data were impossible to acquire due to the lack of critical flow information at the specific site and we are unable to evaluate full flow hydrology levels for these sites. The full hydrology data processing procedure is outlined in a separate document provided by the consultant and is not replicated here. It is critically important for our study that the flow elevations correlate with measured elevations of each vegetation sub-transect. The sub-transect elevations were obtained by ARI staff at the exact locations of the sub-transects, using GPS devices and processing software that was tested and proven sufficiently accurate for the study. While the relative elevation among sub-transect points appears to be accurate, the comparative elevation to the flow levels was in some cases misaligned. This is however a relatively easy fix, once the difference between the transect and flow data elevations is determined, that value is simply added or subtracted from either dataset for calibration. The differences were possible due to water level elevations recorded at each site that could compare flow levels with sub-transect levels directly. A second issue is that the river channels typically decline in elevation from upstream to downstream within a site, resulting in a decline in elevation across the transect locations within a site. This is particularly true for sites further upstream, whereas lowland reaches can be very flat. This means that a single flow level for each site will be inaccurately indicating flow heights at some transect locations. Again, this can be addressed through data calibration for each site so that the flow level speaks to each transect location within a site. This transect-specific calibration is required for all transects within the study with flow data available. This is then a three step process: 1) determine any discrepancy between sub-transect elevations and flow elevations for each site, 2) determine the within-site variation in elevation across transects, 3) provide a calibration value for all transects at all sites with flow data to enable rapid calibration.

Determination of flow events/timelines for evaluation Once accurate and fully aligned data are produced via the processes indicated above, it is important to carefully consider the flow variables to be used within models. Our conceptual understanding of vegetation responses to flows suggests that the depth, duration and

timing (season) are important attributes of flow that dictate responses. These variables can be extracted from the flow data that we have for each site.

Important considerations for flow depth The most important aspect of flow depth is where or not a plant is fully submerged or not by the flow. Our VEFMAP Stage 6 experimental research has shown us that a plant that is only partly submerged is likely to be much less affected than a plant fully submerged. This is the case for most non-aquatic species. Therefore, when we consider the impacts of flow level, the flow elevation is very important, for example, the threshold of plant responses my not occur at the maximum elevation of a flow event but instead at around 40cm below the peak level where most plants will have been fully submerged. However, this impact may vary between species based on their height, growth form, and physiological responses to inundation.

Important considerations for flow duration Flow duration is important because vegetation responses vary dramatically with duration. Our VEFMAP Stage 6 experimental research has shown us that most plants are tolerant of short periods of inundation (and there may be some positive effects at very short durations depending on climate attributes and soil). But as submergence duration increases, most non-aquatic plants will decline in health until they eventually reach a survival threshold. The specific duration of a non-lethal flow may not be critical in evaluating the effect of a flow, but that is critical information for a lethal flow.

Important considerations for flow timing Flow timing is important because vegetation responses vary with flow timing. Our VEFMAP Stage 6 experimental research has shown us that many plants are more negatively affected (more rapidly) by submergence in warmer seasons. Separating winter/spring flow events from summer/autumn events appears to be the most important distinction from this point of view and we don't expect vegetation responses to be equivalent in these periods.

Important considerations for flow regimes When considering flow regimes, we firstly will need to determine a time range that is relevant, e.g. the last 3 years, 5 years,…prior to surveys. We will also probably need to categorise or otherwise quantify waterways with particular regimes to indicate the set of regime components that will influence vegetation patterns. For example, a 'full' regime might have baseflows and/or low-flows as well as at least one spring fresh and summer/autumn fresh. A 'summer' regime might have all but the spring fresh/high-flow, while a 'spring' regime might have all but the summer/autumn fresh. There may also be a category for 'variable' regime within recent years. For categorical regimes, the regime also needs to be consistent across the relevant set of years. Depending on the variability between sites and between years, we might need to consider multiple options for evaluating the regime. Each of the sites needs to be assessed to list the regime within recent years to determine the best options - which has not been done yet.

Determining the variables to use Initially, we will use the most simple flow variables that directly relate to management of e-flows. The most commonly delivered e-flows are baseflows and spring freshes. These events are usually consistently delivered each year at the same flow magnitude. This allows us to evaluate the cumulative effect of those flow levels on vegetation, as well as the effects of a single flow event. This can be done by determining the mean baseflow level and mean peak fresh level from our flow data and using those levels as indicators to compare vegetation trends above and below those elevations. This will be our first point of evaluation for flow. As we progress the analysis and determine how closely or loosely vegetation patterns align with these levels, we will explore alternative approaches and variables for use in the evaluation, such as the number of days inundated by a certain depth over a given period. Importantly, the usefulness of these evaluations depends a lot on the vegetation species or groupings that are evaluated in the models against these flow metrics.

Vegetation categorisation into functional/response groups Different species respond to the environment differently. The same is true for flow responses. This is why all of our vegetation surveys are initially done at the individual species level. However, individual species distributions and abundances are spatially and temporally patchy, so there can be great value in aggregating data from similar species to produce larger data pools. This is the norm for many ecological studies. The critical step is how species are aggregated. There is a balance between too refined groups (with little benefit from increased data pools as the pools increase by small amounts) and too broad groups (there is a large amount of species variation within a group meaning that responses are unclear). One well-established grouping system for wetland plants is by Brock and Casanova (1997) that groups plants by their affinity for different water regimes - called wetland plant functional groups (WPFG). In many cases, the species and the responses are directly equivalent to riparian systems, however there are some notable variations. There are also many different WPFGs, which are impractical to evaluate separately in many cases. In this study, we do not have a pre-determined approach for how vegetation groups will be used. We have compiled all of the relevant WPFG assignments for each species recorded and have determined our own classes in which these groups are nested (with variation required in few cases where the riparian response is expected or observed to be different from the wetland responses). We will initially conduct our analyses using the three hierarchical levels of data aggregation: species < WPFG < broad classes. Based on these initial explorations, we will progress the final evaluation approach. This exploratory approach is important for a study like this where there is not a clear precedent in the published literature for the most appropriate vegetation groupings.

Determination of grazing covariate The most simple form of grazing covariate is a binary score of present or absent. However, as described in Section 2.3, there is a lot of variation in the effect of grazing relating to the intensity (density of animals) and the timing (season of grazing). While we do not have this information, we can roughly categorise the grazing intensity based on site observations, which may be informative. Additionally, we have data on grazer animal (sheep or cattle) which may be important for some or all questions. We need to consider these options and develop candidate variables to test in our models.

Determination of exotic vegetation covariate The exotic vegetation covariate is relatively straightforward in one sense because it will be simple cover and/or diversity estimates that we have collected. However, there is a likely interaction between the impact of exotic species and the terrestrial/riparian grouping, for example, it is possible that terrestrial exotics are less of a problem than riparian ones even with the same amount of cover because they occupy the same habitat and seasonal niches as native riparian species. So this comes back to the vegetation groupings described above which need to be resolved to determine the variables used.

2.5.3 Describe any data exploration or preliminary data analyses

Explanation

In most modelling cases, it is necessary to perform preliminary analyses to understand the data and check that assumptions and requirements of the chosen modelling procedures are met. Data exploration prior to model fitting or development may include exploratory analyses to check for collinearity, spatial and temporal coverage, quality and resolution, outliers, or the need for transformations (Yates et al., 2018).

Preregistration Item

For each separate preliminary or investigatory analysis:

- ☑ State what needs to be known to proceed with further decision-making about the modelling procedure, and why the analysis is necessary.
- ☑ Explain how you will implement this analysis, as well as any techniques you will use to summarise and explore your data.
- ☑ What method will you use to represent this analysis (graphical, tabular, or otherwise, describe)
- ☑ Specify exactly which parts of the data will be used
- ☑ Describe how the results will be interpreted, listing each potential analytic decision, as well as the analysis finding that will trigger each decision, where possible.

Given the complex nature of this analysis and the uncertainties in the data structure due to the data processing that needs to occur, there are many preliminary checks that may need to be conducted. For example, depending on the vegetation groupings that we decide to use, and how we account for 'zero samples' (see Section 2.2) there may be very different amounts of skew or zero-inflation in the response or predictor variables. Initially, we will need to do simple tests of data distributions within hierarchies to check for data spread and prevalence of zeros or outliers. We will also need to assess the collinearity or interactions between candidate predictor variables to ensure relevant interactions are captured and to reduce model overfitting. Our data are spatially and temporally variable with a relatively large spatial coverage but the data density is not equal across space and time, i.e. some sites have been surveyed more than others and in different years. Random effects for site should account for the spatial bias but we will need to carefully consider the interpretation of the outputs and explore options for accounting for temporal variability.

We will conduct a pilot analysis of a subset of the full data-set in order to develop an initial set of candidate models. This pilot data will consist of both hydrological and vegetation data for a single System, *Campaspe*. Although the last three years of data are missing from the current Campaspe dataset, it still contains observations from multiple years which will facilitate setting up a multi-year model. The pilot analysis aims to address uncertainties in how best to specify the candidate models, including how some variables should be operationalised, and which variables should be included given issues like multicolinearity. The analysis aims to identify any major issues that we might need to address within our modelling of the full dataset.

Some questions we will investigate are:

- How do we classify best flow regime based on inundation data?
- What is the relevant regime time-frame to consider for quantifying the regime? 3 / 5 years, something else?
- Are there any outliers?

- What distributional assumptions do we need to make in our models?
- Will there be issues with zero-inflated count data?
- Do the available data support the desired model structures, or do these require simplification?

To resolve these uncertainties we will conduct some exploratory data analyses. Specifically, we will visually assess distributions of the response variables, test correlations between all pairs of numerical (non-categorical) predictor variables, and calculate counts of non-zero observations within each category of any categorical variables included in the analysis (accounted for nested structures and interactions).

We will also fit two initial models to the pilot dataset, one for each key response variable (richness, cover):

```
library(glmmTMB)

# autoregressive model for cover
cover_ar_model <- glmmTMB(
  # specify an autoregressive model structure to model change in plant cover
  plant_hits ~ log_plant_hits_tm1 +
    # assess group- and origin-specific impacts of broad flow "regimes"
    wpfg * origin * (days_above_baseflow + days_above_springfresh) +
    # fixed effects for functional group, zone (bank elev.), period (before/after),
    # and origin (with interactions)
    wpfg * zone * period * origin +
    # fixed effect for grazing impacts (binary variable)
    grazing +
    (1 | site / transect) +   # random effects for transects nested within sites
    (1 | metres) +            # random effect for location of site up the streambank
    (1 | survey_year) +       # random effect for survey year
    family = poisson,         # assume count distribution of plant_hits
    ziformula = ~ wpfg,       # allow zero-inflation, with functional group-specific parameters
    data = veg_cover_ar
)

# model for species richness
richness_model <- glmmTMB(
  species_richness ~
    # assess group- and origin-specific impacts of broad flow "regimes"
    wpfg * origin * (days_above_baseflow + days_above_springfresh) +
    # fixed effects for functional group, zone (bank elev.), period (before/after),
    # and origin (with interactions)
    wpfg * zone * period * origin +
    # fixed effect for grazing impacts (binary variable)
    grazing +
    (1 | site / transect) +   # random effects for transects nested within sites
    (1 | metres) +            # random effect for location of site up the streambank
    (1 | survey_year) +       # random effect for survey year
    offset(npoint),           # offset to account for number of points measured at each transect
    family = poisson,         # assume count distribution of plant_hits
    ziformula = ~ wpfg,       # allow zero-inflation, with functional group-specific parameters
    data = veg_richness
)
```

All models fitted to the pilot data set will undergo the same model checks as the candidate models fitted to the full dataset. The primary focus of model checks for the pilot analysis is to identify appropriate model structures, which has two main steps. First, assessing whether model converged and generated reliable parameter estimates, which in the case of the proposed `glmmTMB` models is assessed internally and printed on model return. Second, assessing whether the model structure (particularly the error distribution) is appropriate for the data, which is supported by posterior predictive checks. The details of these steps are provided in Section 4.3.1.

It is important to note that models fitted to the pilot dataset may not translate directly to the final analysis due to differences in the data structure. These differences will change the distribution of observations within categories, and will introduce a new random effect for water body (or system), which is not required for the single-water body pilot analysis. Due to these changes, the final analysis still requires model checking and may

require changes to the model structure.

2.5.4 Data evaluation, exclusion and missing data

Explanation

Documenting issues with reliability is important because data quality and ecological relevance might be constrained by measurement error, inappropriate experimental design, and heterogeneity and variability inherent in ecological systems (Grimm et al., 2014). Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand (Mahmoud et al., 2009).

Preregistration Item

- ☑ Describe how you will determine how reliable the data are for the given model purpose. Ideally, model input data should be internally consistent across temporal and spatial scales and resolutions, and appropriate to the problem at hand
- ☑ Document any issues with data reliability.
- ☑ How will you determine what data, if any, will be excluded from your analyses?
- ☑ How will outliers be handled? Describe rules for identifying outlier data, and for excluding a site, transect, quadrat, year or season, species, trait, etc.
- ☑ How will you identify and deal with incomplete or missing data?

Data reliability At this stage I am unsure about how we will quantitatively evaluate how reliable the data are for the questions being addressed. Qualitatively, one of the most important considerations will be to assess how logical the outcomes are in relation to our observations and conceptual understanding. We have a very good mental model of these systems due to the large amounts of fieldwork in various areas over the years collecting the dataset used in this study as well as relevant data for other studies. As for data quality, there are likely to be minor errors in such a large dataset that we cannot detect with our quality checking procedures, however, we are very confident that the data have been consistently collected (almost every survey was led by the same individual with the other surveyors being from a consistent set of experienced individuals), in the same places (all transects were permanently marked and relocated at each survey), and the data checking process has been extremely thorough (documented in git). All this considered, we believe that this is one of the most reliable vegetation datasets available anywhere in Victoria for its size and complexity.

Data exclusions Data will only be excluded if we believe it is incorrect, and it can't be corrected, it is irrelevant to a particular model/assessment, or it precludes model fitting. For example, if evaluating survey intervals relating to spring fresh delivery, we can only include years and sites where such a delivery actually occurred. Incorrect data are easily detected where values lie outside possible or plausible ranges, but in other cases they can be very difficult to identify - we are confident that the former have been well accounted for in the datasets but the latter is only partially accounted for and there may be some minor errors that we cannot isolate. All species will be included, but unknown species that have no possible grouping identifier, e.g. native/exotic or life form, may not be possible to include. The vast majority of these occurrences are for seedlings that are too small to be identified. In most cases seedlings had minimal impact on plant cover, so this would have few implications for evaluation of cover, but this may have a greater impact on species richness. Decisions will need to be made for certain unknown species categories, particularly those that are more common in the dataset. At this stage only species that can reliably assigned to a relevant group for a given model will be included. Furthermore, the pilot analysis illustrated extreme zero-inflation for several groups, which prevented model fitting. Consequently these groups were excluded from the pilot study modelling and they may or may not be included in the full dataset analysis, depending on whether the models can be successfully fit.

Plant functional groups `"Atl_native", "Ate_native", "Tda_unknown"` are removed for all models because these `wpfg` are thought to be incorrectly specified (typos).

Missing data There are a small number of cases where we have incomplete or missing data due to various circumstances in the surveys. Currently these are indicated as NA in the dataset and form a small proportion of the overall dataset.

2.6 Conceptual model evaluation

Preregistration Item

- ☑ Describe how your conceptual model will be critically evaluated. Evaluation includes both the completeness and suitability of the overall model structure.
- ☑ How will you critically assess any simplifying assumptions (Augusiak et al., 2014)?
- ☐ Will this process include consultation or feedback from a client, manager, or model user?

Model analysis results will be used to update the existing conceptual model (described in Section 2). Specifically, model results will be used to inform understanding about the short-term (flow models) and long-term (regime models) impacts of environmental flows combined with grazing and exotic vegetation on native vegetation (cover and density) within regulated river channels monitored in this dataset.

3 Formalise and Specify Model

Explanation

In this section describe what quantitative methods you will use to build the model/s, explain how they are relevant to the client/manager/user's purpose.

3.1 Model class, modelling framework and approach

Explanation

Modelling approaches can be described as occurring on a spectrum from correlative or phenomenological to mechanistic or process-based (Yates et al., 2018); where correlative models use mathematical functions fitted to data to describe underlying processes, and mechanistic models explicitly represent processes and details of component parts of a biological system that are expected to give rise to the data (White & Marshall, 2019). A model 'class,' 'family' or 'type' is often used to describe a set of models each of which has a distinct but related sampling distribution (C. C. Liu & Aitkin, 2008). The model family is driven by choices about the types of variables covered and the nature of their treatment, as well as structural features of the model, such as link functions, spatial and temporal scales of processes and their interactions (Jakeman et al., 2006).

Preregistration Item

- ☑ Describe what modelling framework, approach or class of model you will use to implement your model and relate your choice to the model purpose and analytical objectives described in Section 1.1.2 and Section 1.1.3.

We will use a correlative model approach for this analysis, where we will attempt to detect and describe patterns in recorded vegetation data in relation to a series of covariates. Specifically, the analysis aims to provide clear evidence of the influence of environmental flows, and other factors (particularly exotic vegetation and livestock grazing), on native vegetation (cover and diversity) within regulated river channels (Section 1.1.3). We will initially use generalised linear mixed effects models (GLMM) with zero-inflated Poisson family distribution for cover data and species richness (counts). The exact model family will be determined based on an iterative process of model fitting and model checking (with posterior predictive checks) to ensure that the fitted model family is appropriate for the data being modelled (see Section 3.2, below). We believe that GLMMs are a robust approach for our objectives and data structure, with a key feature of the random effects allowing for the hierarchical sampling design and repeated measures.

Although alternative distributions may be suitable for cover data (e.g., binomial, beta), these distributions introduce link functions that complicate the specification of an autoregressive model structure. For this reason, a Poisson distribution was used to model cover data, noting that the Poisson approximates the binomial in the limit of a large number of trials. The Poisson distribution, specified with an offset, still models proportional cover but does not constrain values to sit below an upper bound (values must still be non-negative).

3.2 Choose model features and family

Explanation

All modelling approaches require the selection of model features, which conform with the conceptual model and data specified in previous steps (Jakeman et al., 2006). The choice of model are determined in conjunction with features are selected. Model features include elements such as the functional form of interactions, data structures, measures used to specify links, any bins or discretisation of continuous variables. It is usually difficult to change fundamental features of a model beyond an early stage of model development, so careful thought and planning here is useful to the modeller (Jakeman et al., 2006). However, if changes to these fundamental aspects of the model do need to change, document how and why these choices were made, including any results used to support any changes in the model.

3.2.1 Operationalising Model Variables

Preregistration Item

- ☑ For each response, predictor, and covariate, specify how these variables will be operationalised in the model. This should relate directly to the analytical and/or management objectives specified during the problem formulation phase. Operationalisations could include: the extent of a response, an extreme value, a trend, a long-term mean, a probability distribution, a spatial pattern, a time-series, qualitative change, such as a direction of change or, the frequency, location, or probability of some event occurring. Specify any treatment of model variables, including whether they are lumped / distributed, linear / non-linear, stochastic / deterministic (Jakeman et al., 2006).
- ☑ Provide a rationale for your choices, including why plausible alternatives under consideration were not chosen, and relate your justification back to the purpose, objectives, prior knowledge and or logistical constraints specified in the problem formulation phase (Jakeman et al., 2006).

The two primary response variables are plant cover and species richness (see Section 2.4), which will be operationalised in slightly different ways depending on the time frame:

1. short term-in response to a single event (i.e. before and after); and
2. medium term (2-10 years) in relation to typical flow regimes (flow elevation/duration/timing) over that period.

Both will be using the extent of a response of particular groups of plants at particular bank elevations in relation to flow events (factor condition of before or after an specific event) or flow values (e.g. days of flow to a specific elevation within a year/season), as well as the additional covariates of exotic plant cover and livestock grazing. This will enable an evaluation of variation in plant responses at different elevations in relation to the primary management action (flow) as well as identify the relative impacts of flow and the covariates of exotic plants (interacting with flow) and livestock grazing.

3.2.2 Choose model family

Preregistration Item

- ☑ Specify which family of statistical distributions you will use in your model, and describe any transformations, or link functions.
- ☑ Include in your rational for selection, detail about which variables the model outputs are likely sensitive to, what aspects of their behaviour are important, and any associated spatial or temporal dimensions in sampling.

Posterior predictive checks from the pilot analysis indicated high levels of zero-inflation with some over-dispersion. The final models in the pilot analysis allowed for zero-inflation parameters to differ among plant functional groups (`ziformula = ~ wpfg`) to account for different proportions of zeros among functional groupings, but did not account for over-dispersion. Over-dispersion is commonly accounted for by using negative binomial models, which we could potentially fit using groups-specific dispersion parameters using the `glmmTMB` argument `dispformula = ~ wpfg`. However, all attempts to fit negative binomial models resulted in non-convergence in our pilot analysis. Despite some degree of over-dispersion, given that the zero-inflated Poisson models converged

and reliably captured the proportion of zeros in the pilot dataset, we propose using this approach for all models fitted to the full dataset, including any simplified models.

The two key vegetation response variables have Poisson and binomial families (distributions) as indicated in Section 3.3, but will both be modelled as zero-inflated Poisson distributions for the reasons outlined in Section 3.1. All models will use a log link function. These distributions have been selected based on the type of data and the expected data distributions, particularly the fact that observations of richness and cover are both recorded as counts but include many zero values. There may be unexpected issues with these approaches due to actual data distributions not matching our expectations, such as an unaccounted for over-dispersion of the data.

Should either over-dispersion or zero-inflation be identified in any fitted model, we will try alternative distributions.

Should over dispersion in any model be identified, we will re-fit the models using a negative binomial distribution. However, the models re-fitted using the negative binomial distribution will only be accepted over the Poisson models if they *both* improve over-dispersion *and* do not decline in model performance, particularly model fit, zero-inflation, and posterior predictive checks.

While we are unable to *a priori* precisely weight these criteria in determining the final distribution, the final decision will be guided by the model's overall ability to capture key associations reliably. This is especially likely to occur when there is no 'perfect model' and there is no dominant alternative choice of model distribution.

3.3 Describe *approach* for identifying model structure

Explanation

This section relates to the process of determining the best/most efficient/parsimonious representation of the system at the appropriate scale of concern (Jakeman et al., 2006) that best meets the analytical objectives specified in the problem formulation phase. Model structure refers to the choice of variables included in the model, and the nature of the relationship among those variables. Approaches to finding model structure and parameters may be knowledge-supported, or data-driven (Boets et al., 2015). Model selection methods can include traditional inferential approaches such as unconstrained searches of a dataset for patterns that explain variations in the response variable, or use of ensemble-modelling methods (Barnard et al., 2019). Ensemble modelling procedures might aim to derive a single model, or a multi-model average (Yates et al., 2018). Refining actions to develop a model could include iteratively dropping parameters or adding them, or aggregating / disaggregating system descriptors, such as dimensionality and processes (Jakeman et al., 2006).

Preregistration Item

- ☑ Specify what approach and methods you will use to identify model structure and parameters.
- ☑ If using a knowledge-supported approach to deriving model structure (either in whole or in part), specify model structural features, including:
  - the functional form of interactions (if any)
  - data structures,
  - measures used to specify links,
  - any bins or discretisation of continuous variables (Jakeman et al., 2006),
  - any other relevant features of the model structure.

Structure estimation: model structure specification is knowledge-driven rather than data-driven, with further refinement and simplification guided by the results of the pilot analysis (Section 2.5.3).

Interactions are expected within the proposed models, such as those between elevation and flow, as well as flow, livestock grazing and weeds. For some model options, flow data may not be required, and elevation may be used as a surrogate for flow (e.g. x elevation represents x flow). In this case, a three-way interaction between elevation, grazing and weeds may be used to evaluate vegetation responses. Event-based (short term) models evaluating changes before and after an event would require a fourth interacting term for period (before or after event). Isolating the effects of elevation, flow or period could then be done using post-hoc tests. Data structures are broadly defined in Section 2.2 and Section 2.4 for the different variables proposed for the study. Plant richness data are counts, cover data are hits (successes) from points (trials, where n=40 for all sub-transects) (modelled as *Poisson*, see Section 3.1), flow data may take a range of forms indicated earlier, elevation is provided in *mAHD* but is likely to be input into the model as an ordinal categorical factor with bins based on known elevation of flow

events (e.g. freshes), grazing at this stage may be a binomial indicator of presence or absence, or alternatively as a categorical factor indicating the presence/absence of cattle or sheep.

The model structure will be refined through the pilot analysis (described in Section 2.5.3.), one aspect of which will be to assess model convergence with different levels of interactions (model convergence checks are described in Section 4.3.1). If models do not converge with the above-specified interactions, higher order interactions will be progressively removed from the model (i.e., three way interactions will be removed, then two-way interactions).

The following terms will be included as random effects: transects nested within sites, point location (metres up the streambank), and survey year.

3.4 Describe parameter estimation technique and performance criteria

Explanation

Before calibrating the model to the data, the performance criteria for judging the calibration (or model fit) are specified. These criteria and their underlying assumptions should reflect the desired properties of the parameter estimates / structure (Jakeman et al., 2006). For example, modellers might seek parameter estimates that are robust to outliers, unbiased, and yield appropriate predictive performance. Modellers will need to consider whether the assumptions of the estimation technique yielding those desired properties are suited to the problem at hand. For integrated or sub-divided models, other considerations might include choices about where to disaggregate the model for parameter estimation; e.g. spatial sectioning (streams into reaches) and temporal sectioning (piece-wise linear models) (Jakeman et al., 2006).

3.4.1 Parameter estimation technique

Preregistration Item

- ☑ Specify what technique you will use to estimate parameter values, and how you will supply non-parametric variables and/or data (e.g. distributed boundary conditions). For example, will you calibrate all variables simultaneously by optimising fit of model outputs to observations, or will you parameterise the model in a piecemeal fashion by either direct measurement, inference from secondary data, or some combination (Jakeman et al., 2006).
- ☑ Identify which variables will be parameterised directly, such as by expert elicitation or prior knowledge.
- ☑ Specify which algorithm(s) you will use for any data-driven parameter estimation, including supervised, or unsupervised machine learning, decision-tree, K-nearest neighbour or cluster algorithms (Z. Liu et al., 2018).

Initially, our model structure will be based on our conceptual model of the system/response that has also guided our data collection. In this case, we do not have a very large set of potential environmental variables to select from that may or may not be influential. Instead, we have selected variables and monitoring approaches to collect data that reflect our expectations of correlation/causation. Given this, there is unlikely to be a large need for a model selection process where variables are added or removed sequentially to evaluate relative model performance. However, there will be critical evaluation of all model components to determine their value to the model, such as the hierarchical nature of the data and the need (or not) to include all levels of the hierarchy in the final models. The model selection process is an important part of the study in itself and will be described in the resulting paper, particularly where it relates to relative effects of different variables and relative importance of spatial and temporal scales. Additionally, there will be some investigation of the effectiveness/appropriateness of different forms of certain variables, such as: bank elevation (continuous or categorical), vegetation groupings, and flow variables (Section 2.4). The primary focus of the model is to evaluate the relative impacts of the key variables in line with the conceptual model and hypotheses, so model performance will align with that objective, including model fit parameters of residual plots, model uncertainty, unexplained variation, degrees of freedom, and fit statistics. Evaluating the model performance for predictive capacity within and beyond the dataset may not be required in the current study as it is likely beyond the scope, but this would involve testing predictive capacity within and between waterways using e.g. a cross-validation approach.

Parameterisation: Parameter estimation will be data-driven, and implemented with the `glmmTMB` R package (Brooks et al., 2017), which uses Template Model Builder software to fit flexible GLMM-type models, amongst other model types. `glmmTMB` uses maximum likelihood to estimate parameters for both parametric and non-parametric factor data (nominal or ordinal) variables (e.g. categorised elevation and/or flow data, as well as

binary or categorical grazing variables). We will optimise the model by comparing fit of model outputs to observations (residuals versus fitted values).

3.4.2 Parameter estimation / model fit performance criteria

Preregistration Item

- ☑ Specify which suite of performance criteria you will use to judge the performance of the model. Examples include correlation scores, coefficient of determination, specificity, sensitivity, AUC, etcetera (Yates et al., 2018).
- ☑ Relate any underlying assumptions of each criterion to the desired properties of the model, and justify the choice of performance metric in relation
- ☑ Explain how you will identify which model features or components are significant or meaningful.

R^2, a measure of agreement between fitted and observed values, is the primary performance criterion we will use to evaluate the performance of each model, coupled with posterior predictive checks to assist the suitability of normal structures (Section 4) . When comparing among alternative models, models with a higher R^2 will be preferred over those with lower R^2. As an approximate guide for judging model fit, we will use the following thresholds R^2 in the absence of cross-validation:

- $<$ 0.25 is poor,
- 0.25 - 0.5 is moderate,
- 0.5 - 0.75 is good,
- 0.75 - 1.0 is excellent, but probably indicates overfitting.

3.5 Model assumptions and uncertainties

Preregistration Item

- ☑ Specify assumptions and key uncertainties in the formal model. Describe what gaps exist between the model conception, and the real-world problem, what biases might this introduce and how might this impact any interpretation of the model outputs, and what implications are there for evaluating model-output to inform inferences or decisions?

Poisson and negative binomial models all assume linearity in model parameters, independence between individual observations, as well as the multiplicative effects of independent variables.

Zero-inflation

Based on the results of the pilot study (Section 2.5.3), all Poisson models allow for zero-inflation parameters to differ among plant functional groups (using `glmmTMB::` argument `ziformula=~ wpfg`).

Over-dispersion

Should over-dispersion (where the conditional variance of the outcome variable is greater than the conditional mean) be detected in models specified with a Poisson distribution, negative-binomial models will be fitted because they allow for over-dispersion by estimating the mean and variance independently (Kruppa & Hothorn, 2021) and assume that extra-Poisson variance is a quadratic function of the mean (Lindén & Mäntyniemi, 2011). Negative binomial models may allow for group-specific dispersion parameters using the `glmmTMB::` argument `dispformula = ~ wpfg`.

Issues of over-dispersion are not expected to be fully resolved by the model specifications outlined in Section 3, however proposed model specifications were a compromise between ideal specification and the limitations of our data.

Other Model Assumptions & Sources of Uncertainty

For each specified model we may fit, further assumptions are described below in Section 3.

A formal sensitivity analysis has not yet been completed and so we are unaware of what variables the model outputs are most sensitive to. The many potential forms of flow and/or elevation data is a key area of possible variation in the variable behaviour, which is currently unknown. There are also many spatial and temporal dimensions associated with the data, such as the hierarchical scales of sub-transect, transect, site, reach, waterway, basin, and State. There will be spatial autocorrelation within the data at each of these levels, such as northern

waterways or river basins being more similar to southern counterparts due to climate and geomorphology differences. The extent of the variation among the different spatial scale is currently unknown and is an important aspect of the study to evaluate. Temporal patterns are also important due to the longitudinal nature of the data collection and different seasons of survey (i.e. different proximity to different flow/climate periods). Careful consideration of these aspects will be important in the model design, as per Section 2.4.

3.6 Specify formal model(s)

Explanation

Once critical decisions have been made about the modelling approach and method of model specification, the conceptual model is translated into the quantitative model.

Preregistration Item

- ☑ Specify all formal models
  - ☐ Note, for data-driven approaches to determining model structure and/or parameterisation, it may not be possible to respond to this preregistration item. In such cases, explain why this is the case, and how you will document the model(s) used in the final analysis.
- ☑ For quantitative model selection approaches, including ensemble modelling, specify each model used in the candidate set, including any null or full/global model.

Vegetation Cover Models

Based on the pilot analysis (Section 2.5.3) we have derived three additional model structures to be fitted on the full dataset that are simplified versions of the full models specified in Section 2.5.3. While these simplified models are not ideal, the full models with three-way interactions were too complex given the data, and failed to converge and generate reliable parameter estimates on the pilot dataset. Consequently, the full models will be fitted again to the full dataset.

Full Model, full dataset, GitHub Issue 64

We will attempt to fit a single full model using all data, and adding a random effect for `waterbody` to account for variation between river systems.

Should the full model fit to all data be computationally feasible and converge, we will add the three-way interaction `zone * period * wpfg` back into the full model specification (discarded in the pilot analysis). If this converges it will be used as the basis for final outputs, if not, the simplified models listed below will be used.

Should the full model fitted to the full dataset not be computationally feasible or converge, we will follow a similar strategy for the pilot-study modelling, working from the full model towards simplified models, as there is potential for the full dataset to support a model of intermediate complexity (more complex than the simplified models, but not as complex as the full model). We will fit the following simplified models in the case that the full model fitted to the full dataset does not converge:

Simplified model 1: "flow regime" model, GitHub Issue 65

The aim of this simplified model is to examine how past flow conditions influence vegetation cover while capturing the average effects of flows at different levels over multiple years. The model does not consider zone or period and identifies functional group-specific impacts of days above baseflow or spring fresh levels.

This model includes two flow predictors (`days_above_baseflow_std`, `days_above_spring_fresh_std`), and the full suite of random effects.

```
cover_ar_TMBmod_1 <- glmmTMB::glmmTMB(
   hits ~ log_hits_tm1 +
     days_above_baseflow_std * wpfg * origin +
     days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 + days_above_springfresh_std^2 +
     #   zone * period +
   #  zone + period +
   #  grazing + wpfg  +
     (1 | site / transect) +
     #(1 | site / period) +
```

```
    (1 | metres) +
    (1 | survey_year),
  # offset(npoint),
  family = poisson,
  ziformula=~ wpfg,
  #dispformula =~ wpfg ,
  data = veg_cover_ar_sum |>
    filter(!wpfg_ori %in% c("Atl_native",
                            "Ate_native",
                            "Tda_unknown"))
)
```

Simplified model 2a: "flow events" Model, version 1, GitHub Issue 66

The aim of the 'flow events' models (2a,2b) is to examine how vegetation cover changes in specific zones before and after key flow events (spring and summer freshes). Model 2a seeks to examine the effects on vegetation before and after specific flow events assuming that different plant functional groups (`wpfg`s) have different cover levels but similar responses to flows.

In the pilot analysis, model results indicated that vegetation cover differed in its responses to spring and summer freshes in each zone (below baseflow, baseflow to spring fresh, above spring fresh). Below baseflow level, vegetation increased following the spring fresh and remained high following the summer event (Figure 5, Appendix S7). In the baseflow-to-spring fresh zone, vegetation increased following the spring fresh but returned to pre-spring levels following the summer fresh (Figure 5, Appendix S7). In the above-spring fresh zone, vegetation level declined following both the spring and the summer fresh (Figure 5, Appendix S7).

Consequently, this model includes several categorical predictors as independent fixed effects (`origin`, `wpfg`, `grazing`, and a zone-by-period interaction) and the full suite of random effects, but no `days_above_` predictors.)

This first version of the flow event model allows functional groups to have different levels of cover but assumes that changes following flow events are similar in all groups within a zone. Although this is not an ideal model structure, it is likely that functional groups are restricted to particular zones, in which case the zone-by-period interaction may capture some of the variation attributable to functional groupings.

```
cover_ar_TMBmod_2 <- glmmTMB::glmmTMB(
  hits ~ log_hits_tm1 +
    # days_above_baseflow_std*wpfg*origin +
    # days_above_springfresh_std*wpfg*origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std_sq +
    zone * period +
    origin + wpfg +
    grazing +
    (1 | site / transect) +
    #(1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
  # offset(npoint),
  family = poisson,
  ziformula = ~ wpfg,
  # dispformula =~ wpfg ,
  data = veg_cover_ar_sum |>
    filter(!wpfg_ori %in% c("Atl_native", "Ate_native", "Tda_unknown"))
)
```

Simplified model 2b: "flow events" model, version 2, GitHub Issue 67

The aim is to examine how vegetation cover of each functional grouping changes before and after key flow events (spring and summer freshes). This second version (Simplified model 2b) allows functional groups to have different responses to flow events (spring and summer freshes) but assumes that vegetation in all zones changes similarly following each flow event. As for simplified model 2a, this model structure is not ideal, but provides a method to distinguish `wpfg`-specific responses to flow events.

This model includes several categorical predictors as independent fixed effects (`origin`, `zone`, `grazing`), a `wpfg`-by-`period` interaction, as well as the full suite of random effects, and no `days_above_` predictors.)

```
cover_ar_TMBmod_3 <- glmmTMB::glmmTMB(
  hits ~ log_hits_tm1 +
    # days_above_baseflow_std*wpfg*origin +
    # days_above_springfresh_std*wpfg*origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std_sq +
    zone + wpfg * period +
    origin +
    grazing +
    (1 | site / transect) +
    # (1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
  # offset(npoint),
  family = poisson,
  ziformula = ~ wpfg,
  # dispformula =~ wpfg ,
  data = veg_cover_ar_sum |>
    filter(!wpfg_ori %in% c("Atl_native", "Ate_native", "Tda_unknown"))
)
```

Species Richness Models

Full Model, Full Dataset, GitHub Issue 68

Simplified Model 1: "Flow Regime" model, GitHub Issue 69

```
richness_ar_TMBmod_1 <- glmmTMB::glmmTMB(
  richness ~
    days_above_baseflow_std * wpfg * origin +
    days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std^2 +
    # zone * period +
    #zone *period + zone*wpfg + wpfg*period +
    # grazing + origin +
    (1 | site / transect) +
    #(1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
  # offset(npoint),
  family = poisson,
  #family = nbinom2,
  #ziformula=~ wpfg,
  # dispformula =~ wpfg ,
  data = veg_richness |>
    filter(!wpfg_ori %in% c("Atl_native",
                            "Ate_native",
                            "Tda_unknown"))
)
```

Simplified Model 2: "Flow Events" model, GitHub Issue 70

Rather than splitting the flow events model into two different versions, as we did for the cover models, we have combined these into a single model for richness.

```
richness_ar_TMBmod_2 <- glmmTMB::glmmTMB(
  richness ~
    #days_above_baseflow_std +
    # days_above_springfresh_std +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std^2 +
    # zone * period +
    zone * period + zone * wpfg + wpfg * period +
```

```
    grazing + origin +
    (1 | site / transect) +
    # (1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
  # offset(npoint),
  family = poisson,
  #family = nbinom2,
  #ziformula=~ wpfg,
  # dispformula =~ wpfg ,
  data = veg_richness |>
    filter(!wpfg_ori %in% c("Atl_native",
                            "Ate_native",
                            "Tda_unknown")) |>
    filter(!wpfg %in% c("Sk", "Se"))
)
```

4 Model Calibration, Validation & Checking

4.1 Model calibration and validation scheme

Explanation

This section pertains to any data calibration, validation or testing schemes that will be implemented. For example, the model may be tested on data independent of those used to parameterise the model (external validation), or the model may be cross-validated on random sub-samples of the data used to parameterise the model (internal cross-validation Yates et al., 2018; Barnard et al., 2019). For some types of models, hyper-parameters are estimated from data, and may be tuned on further independent holdouts of the training data ("validation data").

Preregistration Item

☐ Describe any data calibration, validation and testing scheme you will implement, including any procedures for tuning or estimating model hyper-parameters (if any).

4.1.1 Describe calibration/validation data

Explanation & Rationale

The following items pertain to properties of the *datasets* used for calibration (training), validation, and testing.

Preregistration Item

If using external / independent holdout data for model testing and evaluation (delete as needed):

☑ Which data will be used as the testing data? What method will you be used for generating training / test data subsets?

☑ Describe any known differences between the training/validation and testing datasets, the relative size of each, as well as any stratification methods used for ensuring evenness of groups between data sets.

☑ It is preferable that any independent data used for model testing remains unknown to modellers during the process of model development. Describe the relationship modellers have with model validation data. Will independent datasets be known or accessible to any modeller or analyst?

Due to the complexity of the analysis, and the focus on inference, no data partitioning / testing on external data will be used. However, we will use a subset of the data for exploratory pilot analysis that informs the final model specifications (Campaspe catchment only, Section 2.5.3).

4.2 Implementation verification

Explanation & Examples

Model implementation verification is the process of ensuring that the model has been correctly implemented, and that the model performs as described by the model description (Grimm et al., 2014). This process is distinct from model checking, which assesses the model's performance in representing the system of interest (Conn et al., 2018).

- Checks for verification implementation should include i) thoroughly checking for bugs or programming errors, and ii) whether the implemented model performs as described by the model description (Grimm et al., 2014).
- Qualitative tests could include syntax checking of code, and peer-code review (Ivimey et al., 2023). Technical measures include using unit tests, or in-built checks within functions to prevent potential errors.

Preregistration Item

☑ What Quality Assurance measures will you take to verify the model has been correctly implemented? Specifying a priori quality assurance tests for implementation verification may help to avoid selective debugging and silent errors.

Implementation verification will be assessed using a number of techniques, but will broadly follow the approach proposed by Ivimey-Cook *et al.* (2023):

1. Code will be reviewed periodically using the GitHub flow model where code is submitted for independent review by collaborators before being merged into the working copy of the code repository. Code will be assessed either by attempting to reproduce the code in the pull-request, or by visual inspection.
2. Defensive programming techniques, in-line error checking, functionalisation, modularisation and documentation of analysis code will be used as preventative measures to catch bugs and ensure proper code implementation (Ivimey et al., 2023),
3. Finally, the final analysis and results will be subjected to a more substantial peer-code review from collaborators further removed from code writing and analysis implementation (likely CJ or EG), following the 4R's (Ivimey et al., 2023).

4.3 Model checking

Rationale & Explanation

"Model Checking" goes by many names ("conditional verification", "quantitative verification", "model output verification"), and refers to a series of analyses that assess a model's performance in representing the system of interest (Conn et al., 2018). Model checking aids in diagnosing assumption violations, and reveals where a model might need to be altered to better represent the data, and therefore system (Conn et al., 2018). Quantitative model checking diagnostics include goodness of fit, tests on residuals or errors, such as for heteroscedascity, cross-correlation, and autocorrelation (Jakeman et al., 2006).

4.3.1 Quantitative model checking

Preregistration Item

During this process, observed data, or data and patterns that guided model design and calibration, are compared to model output in order to identify if and where there are any systematic differences.

☑ Specify any diagnostics or tests you will use during model checking to assess a model's performance in representing the system of interest.

☑ For each test, specify the criteria that will you use to interpret the outcome of the test in assessing the model's ability to sufficiently represent the gathered data used to develop and parameterise the model.

Posterior predictive checks will be conducted to assess the degree of zero-inflation and over-dispersion. Posterior checks compare the distribution of observed data against a distribution simulated from the fitted model and

are assessed quantitatively or graphically (Conn et al., 2018). We will plot both distributions and visually compare the distributions: close alignment of the two distributions indicates that the specified model structure is appropriate, whereas deviations between the observed distribution and the model-generated distribution indicate potential assumption violations and mismatches between the data and model.

For graphical posterior checks, there is no threshold but any disagreement, particularly in key aspects of the model (zero inflation, and counts in the range of the majority of the data) would suggest an alternative distribution is worth considering. Often no distribution will be perfect, so it's a choice between two imperfect options. (in which case, we will take into consideration the balance between model balance simplicity, R^2, and the "key aspects" of model such as zero inflation and the bulk of the data).

Should over-dispersion and zero-inflation be present in the fitted models, alternative model families / structures (Section 3.2) may be investigated and compared to the primary models outlined in Section 2.5.3.

4.3.2 Qualitative model checking

ℹ Explanation

This step is largely informal and case-specific, but requires, 'face validation' with model users / clients / managers who aren't involved in the development of the model to assess whether the interactions and outcomes of the model are feasible an defensible (Grimm et al., 2014). This process is sometimes called a "laugh test" or a "pub test" and in addition to checking the model's believability, it builds the client's confidence in the model (Jakeman et al., 2006). Face validation could include structured walk-throughs, or presenting descriptions, visualisations or summaries of model results to experts for assessment.

🔥 Preregistration Item

- ☑ Briefly explain how you will qualitatively check the model, and whether and how you will include users and clients in the process.

In terms of qualitative assessment, model results will be checked by field experts Chris Jones & Lyndsey Vivian, who will assess model-estimated associations for their plausibility given their expert knowledge of the underlying target system.

4.3.3 Assumption Violation Checks

🔥 Preregistration Item

The consequences of assumption violations on the interpretation of results should be assessed (Araújo et al., 2019).

- ☐ Explain how you will demonstrate robustness to model assumptions and check for violations of model assumptions.
- ☐ If you cannot perform quantitative assumption checks, describe what theoretical justifications would justify a lack of violation of or robustness to model assumptions.
- ☐ If you cannot demonstrate or theoretically justify violation or robustness to assumptions, explain why not, and specify whether you will discuss assumption violations and their consequences for interpretation of model outputs.
- ☐ If assumption violations cannot be avoided, explain how you will explore the consequences of assumption violations on the interpretation of results. *This step is to be completed in interim iterations of the preregistration, only if there are departures from assumptions as demonstrated in the planned tests above.*

5 Model Validation and Evaluation

Explanation

The model validation & evaluation phase comprises a suite of analyses that collectively inform inferences about whether, and under what conditions, a model is suitable to meet its intended purpose (Augusiak et al., 2014). Errors in design and implementation of the model and their implication on the model output are assessed. Ideally independent data are used against the model outputs to assess whether the model output behaviour exhibits the required accuracy for the model's intended purpose. The outcomes of these analyses build confidence in the model applications and increase understanding of model strengths and limitations. Model evaluation, including model analysis, should complement model checking. It should evaluate model checking, and consider over-fitting and extrapolation. As the proportion of calibrated or uncertain parameters increases, so does the risk that the model seemingly works correctly, but for the wrong mechanistic reasons (Boettiger, 2022). Evaluation thus complements model checking because we can rule out the chance that the model fits the calibration data well, but has not captured the relevant ecological mechanisms of the system pertinent to the research question or the decision problem underpinning the model (Grimm et al., 2014). Evaluation of model outputs against external data in conjunction with the results from model checking provide information about the structural realism and therefore credibility of the model (Grimm & Berger, 2016).

5.1 Model output corroboration

Explanation

Ideally, model outputs or predictions are compared to independent data and patterns that were not used to develop, parameterise, or verify the model. Testing against a dataset of response and predictor variables that are spatially and/or temporally independent from the training dataset minimises the risk of artificially inflating model performance measures (Araújo et al., 2019). Although the corroboration of model outputs against an independent validation dataset is considered the 'gold standard' for showing that a model properly represents the internal organisation of the system, model validation is not always possible because empirical experiments are infeasible or model users are working on rapid-response time-frames, hence, why ecologists often model in the first place (Grimm et al., 2014). Independent predictions might instead be tested on sub-models. Alternatively, patterns in model output that are robust and seem characteristic of the system can be identified and evaluated in consultation with the literature or by experts to judge how accurate the model output is (Grimm et al., 2014).

Preregistration Item

- ☑ State whether you will corroborate the model outputs on external data, and document any independent validation data in this step.
- ☑ It is preferable that any independent data used for model evaluation remains unknown to modellers during the process of model building (Dwork et al., 2015), describe the relationship modellers have to model validation data, e.g. will independent datasets be known to any modeller or analyst involved in the model building process?
- ☑ If unable to evaluate the model outputs against independent data, explain why and explain what steps you will take to interrogate the model.

Model evaluation and validation on external data will not be undertaken in this study due to the complexity of the models (hierarchical structure combined with interactions) and data (potential for over-dispersion and zero-inflation) and also due to the focus of the analysis on inference.

For the same reasons, we will not partition the data for model validation on independent data using cross-validation or other data-partitioning approaches. Instead, in-sample model fit assessment will be conducted (Section 3.4.2).

Models may be validated on newly collected independent data in the future, but external validation is outside the scope of this study.

We have partitioned the data for pilot analysis, see Section 2.5.3 for details of this subsetting. Pilot analysis was

performed blind to the full dataset, and was performed only on the pilot data subset.

5.2 Choose performance metrics and criteria

Explanation

Model performance can be quantified by a range of tests, including measures of agreement between predictions and independent observations, or estimates of accuracy, bias, calibration, discrimination refinement, resolution and skill (Araújo et al., 2019). Note that the performance metrics and criteria in this section are used for evaluating the structured and parameterised models (ideally) on independent holdout data, so this step is additional to any performance criteria used for determining model structure or parameterisation (Section 3.4.2).

Preregistration Item

- ☐ Specify what performance measures you will use to evaluate the model and briefly explain how each test relates to different desired properties of a model's performance.
- ☐ Spatial, temporal and environmental pattern of errors and variance can change the interpretation of model predictions and conservation decisions (Araújo et al., 2019). Where relevant and possible, describe how you will characterise and report the spatial, temporal and environmental pattern of errors and variance.
- ☐ If comparing alternative models, specify what measures of model comparison or out-of-sample performance metrics will you use to find support for alternative models or else to optimise predictive ability. State what numerical threshold or qualities you will use for each of these metrics.

5.3 Model analysis

Rationale & Explanation

Uncertainty in models arises due to incomplete system understanding (which processes to include, or which interact), from imprecise, finite and sparse data measurements, and from uncertainty in input conditions and scenarios for model simulations or runs (Jakeman et al., 2006). Non-technical uncertainties can also be introduced throughout the modelling process, such as uncertainties arising from issues in problem-framing, indeterminicies, and modeller / client values (Jakeman et al., 2006).

The purpose of model analysis is to prevent blind trust in the model by understanding how model outputs have emerged, and to 'challenge' the model by verifying whether the model is still believable and fit for purpose if one or more parameters are changed (Grimm et al., 2014).

Model analysis should increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system. Model analysis typically consists of sensitivity analyses preceded by uncertainty analyses (Saltelli et al., 2019), and a suite of other simulation or other computational experiments. The aim of such computational experiments is to increase understanding of the model behaviour by identifying which processes and process interactions explain characteristic behaviours of the model system (Grimm et al., 2014). Uncertainty analyses and sensitivity analyses augment one another to draw conclusions about model uncertainty.

Because the results from a full suite of sensitivity analysis and uncertainty analysis can be difficult to interpret due to the number and complexity of causal relations examined (Jakeman et al., 2006), it is useful for the analyst to relate the choice of analysis to the modelling context, purpose and analytical objectives defined in the problem formulation phase, in tandem with any critical uncertainties that have emerged during model development and testing prior to this point.

5.3.1 Uncertainty Analyses

Explanation

Uncertainty can arise from different modelling techniques, response data and predictor variables (Araújo et al., 2019). Uncertainty analyses characterise the uncertainty in model outputs, and identify how uncertainty in model parameters affects uncertainty in model output, but does not identify which model assumptions are driving this behaviour (Grimm et al., 2014; Saltelli et al., 2019). Uncertainty analyses can include propagating known uncertainties through the model, or by investigating the effect of different model scenarios with different parameters and modelling technique (Araújo et al., 2019), for example. It could also include characterising the output distribution, such as through empirical construction using model output data points. It could also include extracting summary statistics like the mean, median and variance from this distribution, and perhaps constructing confidence intervals on the mean

Preregistration Item

- ☐ Please describe how you will characterise model and data uncertainties, e.g. propagating known uncertainties through the model, investigating the effect of different model scenarios with different parameters and modelling technique combinations (Araújo et al., 2019), or empirically constructing model distributions from model output data points, and extracting summary statistics, including the mean, median, variance, and constructing confidence intervals (Saltelli et al., 2019).
- ☐ Relate your choice of analysis to the context and purposes of the model described in the problem formulation phase. For instance, discrepancies between model output and observed output may be important for forecasting models, where cost, benefit, an risk over a substantial period must be gauged, but much less critical for decision-making or management models where the user may be satisfied with knowing that the predicted ranking order of impacts of alternative scenarios or management options is likely to be correct, with only a rough indication of their sizes" (Jakeman et al., 2006).
- ☐ Briefly describe how you will summarise the results of these in silico experiments with graphical, tabular, or other devices, such as summary statistics.
- ☐ If the chosen modelling approach is able to explicitly articulate uncertainty due to data, measurements or baseline conditions, such as by providing estimates of uncertainty (typically in the form of probabilistic parameter covariance, Jakeman et al., 2006), specify which measure of uncertainty you will use.

Uncertainty analyses will not be conducted.

5.3.2 Sensitivity analyses

Explanation

Sensitivity analysis examines how uncertainty in model outputs can be apportioned to different sources of uncertainty in model input (Saltelli et al., 2019).

Preregistration Item

- ☐ Describe the sensitivity analysis approach you will take: deterministic sensitivity, stochastic sensitivity (variability in the model), or scenario sensitivity (effect of changes based on scenarios).
- ☐ Describe any sensitivity analyses you will conduct by specifying which parameters will be held constant, which will be varied, and the range and intervals of values over which those parameters will be varied.
- ☐ State the primary objective of each sensitivity analysis, for example, to identify which input variables contribute the most to model uncertainty so that these variables can be targeted for further data collection, or alternatively to identify which variables or factors contribute little to overall model outputs, and so can be 'dropped' from future iterations of the model (Saltelli et al., 2019).

No sensitivity analyses will be performed on any models developed in this study.

5.3.3 Model application or scenario analysis

Preregistration Item

- ☐ Specify any input conditions and relevant parameter values for initial environmental conditions and decision-variables under each scenario specified in Section 1.
- ☐ Describe any other relevant technical details of model application, such as methods for how you will implement any simulations or model projections.
- ☑ What raw and transformed model outputs will you extract from the model simulations or projections, and how will you map, plot, or otherwise display and synthesise the results of scenario and model analyses.
- ☐ Explain how you will analyse the outputs to answer your analytical objectives. For instance, describe any trade-off or robustness analyses you will undertake to help evaluate and choose between different alternatives in consultation with experts or decision-makers.

Model analysis

Model features will be identified using visual plots of estimates, i.e. forest plots of most coefficients, and marginal effects plots (plots of predicted values of y against values of x) under different combinations of of the key flow, zone, and period predictor variables.

Forest Plots: For each fitted model we will extract both fixed-effect and random-effect coefficient estimates and their 95% Confidence Intervals using `parameters::model_parameters()` (Lüdecke et al., 2020) or similar. Estimates will then be visualised as forest plots in R.

Predicted Values Plots: We will use the `effects::effect_plot()` function (Fox & Hong, 2009), or similar, to create marginal effects plots for each fitted model.

Scenario analysis will not be used in this study, but will be used in future predictive modelling studies based on the models developed in this study.

5.3.4 Other simulation experiments / robustness analyses

Preregistration Item

- ☐ Describe any other simulation experiments, robustness analyses or other analyses you will perform on the model, including any metrics and their criteria / thresholds for interpreting the results of the analysis.

No further simulation experiments, robustness or other analyses will be performed on the model other than analyses described above in this document.

References

Araújo, M., Anderson, R., Márcia Barbosa, A., Beale, C., Dormann, C., Early, R., Garcia, R., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R., Zimmermann, N., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Sci Adv*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Barnard, D. M., Germino, M. J., Pilliod, D. S., Arkle, R. S., Applestein, C., Davidson, B. E., & Fisk, M. R. (2019). Can't see the random forest for the decision trees: Selecting predictive models for restoration ecology. *Restoration Ecology*. https://doi.org/10.1111/rec.12938

Boets, P., Landuyt, D., Everaert, G., Broekx, S., & Goethals, P. L. M. (2015). *Evaluation and comparison of data-driven and knowledge-supported bayesian belief networks to assess the habitat suitability for alien macroinvertebrates*. *74*, 92–103. https://doi.org/10.1016/j.envsoft.2015.09.005

Boettiger, C. (2022). The forecast trap. *Ecol Lett*. https://doi.org/10.1111/ele.14024

Brock, M. A., & Casanova, M. T. (1997). *Frontiers in ecology; building the links* (N. Klomp & I. Lunt, Eds.; pp. 181–192). Elsevier Science.

Brooks, M. E., Kristensen, K., Benthem, J. ,. van, Koen, Magnusson, A., Berg, Casper,W., Nielsen, A., Skaug, Hans,J., Mächler, M., & Bolker, Benjamin,M. (2017). glmmTMB balances speed and flexibility

among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378. https://doi.org/10.32614/RJ-2017-066

Cartwright, S. J., Bowgen, K. M., Collop, C., Hyder, K., Nabe-Nielsen, J., Stafford, R., Stillman, R. A., Thorpe, R. B., & Sibly, R. M. (2016). Communicating complex ecological models to non-scientist end users. *Ecological Modelling*, *338*, 51–59. https://doi.org/10.1016/j.ecolmodel.2016.07.012

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, *9*, 341–317. https://doi.org/10.1002/ecm.1314

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Fox, J., & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, *32*(1), 1–24. https://doi.org/10.18637/jss.v032.i01

Fraser, H., Rumpff, L., Yen, J. D. L., Robinson, D., & Wintle, B. A. (2017). Integrated models to support multiobjective ecological restoration decisions. *Conservation Biology*, *31*(6), 1418–1427. https://doi.org/10.1111/cobi.12939

Grimm, V., Augusiak, J., Focks, A., Frank, B. M., Gabsi, F., Johnston, A. S. A., Liu, C., Martin, B. T., Meli, M., Radchuk, V., Thorbek, P., & Railsback, S. F. (2014). Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, *280*, 129–139. https://doi.org/10.1016/j.ecolmodel.2014.01.018

Grimm, V., & Berger, U. (2016). Structural realism, emergence, and predictions in next-generation ecological modelling: Synthesis from a special issue. *Ecological Modelling*, *326*, 177–187. https://doi.org/10.1016/j.ecolmodel.2016.01.001

Ivimey, E. R., Pick, J. L., Bairos, K. R., Culina, A., Gould, E., Grainger, M., Marshall, B. M., Moreau, D., Paquet, M., Royauté, R., Sánchez, A., Silva, I., & Windecker, S. M. (2023). Implementing code review in the scientific workflow: Insights from ecology and evolutionary biology. *Journal of Evolutionary Biology*. https://doi.org/https://doi.org/10.1111/jeb.14230

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Kruppa, J., & Hothorn, L. (2021). A comparison study on modeling of clustered and overdispersed count data for multiple comparisons. *Journal of Applied Statistics*, *48*(16), 3220–3232. https://doi.org/10.1080/02664763.2020.1788518

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414–1421. https://doi.org/10.1890/10-1831.1

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002

Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., & Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: Recent progress and future challenges. *Environmental Reviews*, *26*(4), 339–350. https://doi.org/https://doi.org/10.1139/er-2018-0034

Lüdecke, D., Ben-Shachar, M., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using r. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Mahmoud, M., Liu, Y., Hartmann, H., Stewart, S., Wagener, T., Semmens, D., Stewart, R., Gupta, H., Dominguez, D., Dominguez, F., Hulse, D., Letcher, R., Rashleigh, B., Smith, C., Street, R., Ticehurst, J., Twery, M., Delden, H. van, Waldick, R., … Winter, L. (2009). A formal framework for scenario development in support of environmental decision-making. *Environmental Modelling & Software*, *24*(7), 798–808. https://doi.org/10.1016/j.envsoft.2008.11.010

McCarthy, M. A., Thompson, C. J., Moore, A. L., & Possingham, H. P. (2011). Designing nature reserves in the face of uncertainty. *Ecology Letters*, *14*(5), 470–475. https://doi.org/10.1111/j.1461-0248.2011.01608.x

McDonald-Madden, E., Baxter, P. W. J., & Possingham, H. P. (2008). Making robust decisions for conservation with restricted money and knowledge. *Journal of Applied Ecology*, *45*(6), 1630–1638. https://doi.org/10.1111/j.1365-2664.2008.01553.x

Moallemi, E. A., Elsawah, S., & Ryan, M. J. (2019). Strengthening "good" modelling practices in robust decision support: A reporting guideline for combining multiple model-based methods. *Mathematics and Computers in Simulation*. https://doi.org/10.1016/j.matcom.2019.05.002

Moon, K., Guerrero, A. M., Adams, Vanessa. M., Biggs, D., Blackman, D. A., Craven, L., Dickinson, H., & Ross, H. (2019). Mental models for conservation research and practice. *Conservation Letters*, *12*(3), e12642. https://doi.org/10.1111/conl.12642

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., Li, S., & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, *114*, 29–39. https://doi.org/10.1016/j.envsoft.2019.01.012

Tonkin, Z., Jones, C., Clunie, P., Vivian, L., Amtstaetter, F., Jones, M., Koster, W., Mole, B., O'Connor, J.,

Brooks, J., Caffrey, L., & Lyon, J. (2020). *Victorian environmental flows monitoring and assessment program.*

White, C. R., & Marshall, D. J. (2019). Should we care if models are phenomenological or mechanistic. *Trends in Ecology & Evolution*, *34*(4), 276–278. https://doi.org/10.1016/j.tree.2019.01.006

Yates, K., Bouchet, P., Caley, M., Mengersen, K., Randin, C., Parnell, S., Fielding, A., Bamford, A., Ban, S., Barbosa, A., Dormann, C., Elith, J., Embling, C., Ervin, G., Fisher, R., Gould, S., Graf, R., Gregr, E., Halpin, P., … Sequeira, A. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol. (Amst.)*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

D.6 Case Study Pilot Analysis Summary Report

D.6.1 Pilot Analysis of VEFMAP Riparian Vegetation Surveys

Date: 2024-01-24 Authors: Henry Wootton & Jian Yen

A pilot study was conducted to identify model structures that were estimable (i.e., generated reliable parameter estimates) and captured key features of the data (e.g., many zeros). The details are outlined in the preregistration document. In short, the process began with a preferred model structure and refined this by removing interactions until a suitable model was identified.

The model for vegetation cover included an additional term for the previous cover in the previous survey, which made this an autoregressive model (i.e., modelling the change in cover rather than absolute cover).

The flow predictors were simple metrics intended to capture broad features of the flow regime at a site. The two metrics were the number of days in which discharge was above baseflow levels, and the number of days in which discharge was above spring fresh levels. Values were calculated for each water year, and a 3-year rolling mean was used to represent broad changes in the multi-year flow regime (cf. variation within or among individual years). Quadratic transforms of these metrics were also included in the model.

Models for cover and species richness were fitted with zero-inflated Poisson distributions, with negative binomial distributions as an alternative to address over-dispersion. For cover, this differs from the (generally preferred) use of a bounded distribution (e.g., binomial or beta). However, parameters of bounded distributions are challenging to estimate when many values are near the bounds (lower or upper), and typically introduce link functions that do not allow straightforward specifications of autoregressive models. For this reason,

```
Veg. response ~ wpfg * origin * flow predictors +
 wpfg * origin * zone * period +
 grazing +
 (1 | site / transect) +
 (1 | transect / metres) +
 (1 | survey_year) +
 (1 | species) +
 offset(npoints)
```

Listing D.1: The preferred full model structure (in high-level, R notation).

the Poisson distribution was used in models of cover, noting that this approximates the binomial distribution in the limit of large numbers of trials.

Initial models aimed to include data at the level of species. This was infeasible for two reasons. First, data resolved to the level of species required an extremely large data set (due to the many species recorded), which made model estimation computationally challenging. Second, the large data set (with all species) comprised mostly zero observations, resulting in extreme zero-inflation that affected model convergence. The final models were fitted to cover summed over all species within a plant functional grouping, noting that these summed cover estimates sometimes (but rarely) exceeded 100% due to species overlap.

Key decisions and final model structures

1. High-level interactions are too challenging to estimate. Instead proposed three models for vegetation cover:
    1. A flow regime model that examines flow predictors only (with random effects).
    2. A flow event model (version 1) that examines `zone`-by-`period` interactions but assumes that all functional groups have similar patterns (within a `zone` and `period`).
    3. A flow event model (version 2) that examines functional group-by-`period` interactions but assumes that all zones respond similarly.
2. Species richness model is still being finalised, but should at least be able to encompass versions 1 and 2 of the cover model in a single analysis.
3. These models ask high-level questions about the effects of flows (and flow events) on vegetation cover and species richness and provide generalisable insight into how flows could be delivered to influence vegetation outcomes.
4. Quadratic flow effects too complex to fit: removed.
5. Metres within transect too complex to include: replaced with a random effect for metres, which assumes that the distance up the bank might influence vegetation outcomes but in similar ways among sites.
6. Zero-inflation specified as a function of plant functional group to allow for different proportions of zeros among groupings.
7. Over-dispersion present but not resolved by negative binomial model (even with `wpfg`-specific dispersion parameters): zero-inflated Poisson model retained, acknowledging lingering issues of over-dispersion.

8. Offset removed because all observations were based on the same number of points.

Decision points for full analysis

Several decisions will need to be made prior to the full analysis (which should begin ASAP).

1. Assess whether the full `zone`-by-`period`-by- `wpfg` interaction is possible with the full data set. If not, use simplified models outlined above.
2. Decide whether to fit a single model to all waterbodies or separate models for each (decision based on model convergence and posterior checks – a single model is preferable if feasible).
3. Decide on how to visualise model outputs.

D.6.2 Vegetation Cover: regime model

This model includes two flow predictors (days above baseflow, days above spring fresh) and the full suite of random effects. The aim is to examine how past flow conditions influence vegetation cover. The model does not consider `zone` or `period`.

The model identifies functional group-specific impacts of days above baseflow or spring fresh levels. Interpretations here are based on the predicted associations (Figure D.2 and Figure D.3) because interpreting forest plots of coefficients (Figure D.1) is challenging in models with many interactions. Positive effects of baseflows are clearest for native `ATl` and `ATe` and exotic `Tdr` and `Tda`, all of which have higher cover under baseflows. Positive effects of spring freshes are clearest for native `Sk`, `Tda`, and `Ate` and exotic `ATe`. Negative effects of spring freshes are clearest for `Tdr.`

Table D.1: *Vegetation cover flow regime model summary. Signif. codes: 0 '∗∗∗' 0.001 '∗∗' 0.01 '∗' 0.05 '.' 0.1 ' ' 1.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
(Intercept)	-6.338926	3.945021	-1.61	0.1081
log_hits_tml	0.466865	0.084717	98.98	< 2e-16 ∗∗∗
days_above_baseflow_std	3.208391	3.369341	0.95	0.3410
wpgATe	6.158428	3.941779	1.56	0.1182
wpfgATl	3.547942	0.684107		5.19 2.15e-07 ∗∗∗
wpfgSe	-21.385108	25.954264	-0.82	0.4100
wpfgsk	1.108790	1.345171	0.82	0.4132
wpfgTda	7.307450	3.941643	1.85	0.0638
wp fgTdr	7.736796	3.941596	1.96	0.0497 ∗
originative	2.784402	3.994830	0.70	0.4858

Table D.1: *Vegetation cover flow regime model summary. Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
days_above_springfresh_std	-0.123400	0.390857	-0.32	0.7522
days_above_base-flow_std:wpfgATe	-2.987176	3.367936	-0.89	0.3751
days_above_base-flow_std:wpfgATL	0.155333	0.657344	0.24	0.8132
days_above_base-flow_std:wpfgSe	14.468021	26.401090	0.55	0.5837
days_above_base-flow_std:wpfgSk	0.182377	1.265110	0.14	0.8854
days_above_base-flow_std:wpfgTda	-2.975603	3.367609	-0.88	0.3769
days_above_base-flow_std:wpfgTdr	-2.989015	3.367573	-0.89	0.3748
days_above_base-flow_std:originnative	-2.304494	3.430569	-0.67	0.5017
wpfgAle: originnative	-1. 514880	3.995046	-0.38	0.7045
wpfgATl:originnative	NA	NA	NA	NA
wpfgSe: originnative	NA	NA	NA	NA
wpfgSk:originnative	NA	NA	NA	NA
wpfgida: originnative	-2.579748	3.994900	-0.65	0.5184
wpfgTdr: originnative	-3.710273	3.995021	-0.93	0.3530
wpf-gATe:days_above_springfresh_std	1.010669	0.382893	2.64	0.0083 **
wpf-gATl:days_above_springfresh_std	-1.755976	0.272420		-6.45 1.15c-10 ***
wpfgSe:days_above_springfresh_std	2.328011	3.084072	0.75	0.4503
wpfgSk:days_above_springfresh_std	-0.323750	0.546548	-0.59	0.5536
wpfgTda:days_above_springfresh_std	0.255564	0.381392	0.67	0.5028
wpfgTdr:days_above_springfresh_std	0.271463	0.381315	0.71	0.4765
originna-tive:days_above_springfresh_std	2.036161	0.468210		4.35 1.37e-05 ***
days_above_base-flow_std:wpfgATe:originnative	2.376384	3.430972	0.69	0.4885
days_above_base-flow_std:wpfgATl:originnative	NA	NA	NA	NA
days_above_base-flow_std:wpfgSe:originnative	NA	NA	NA	NA
days_above_base-flow_std:wpfg5k:originnative	NA	NA	NA	NA
days_above_base-flow_std:wpfgTda:originnative	2.166421	3.430644	0.63	0.5277
days_above_base-flow_std:wpfgTdr:originnative	2.291317	3.430663	0.67	0.5042
wpfgATe:originna-tive:days_above_springfresh_std	-2.761693	0.469796	-5.88	4.14e-09 ***
wpfgATl:originna-tive:days_above_springfresh_std	NA	NA	NA	NA
wpfgSe:originna-tive:days_above_springfresh_std	NA	NA	NA	NA
wpfgSk:originna-tive:days_above_springfresh_std	NA	NA	NA	NA
wpfgTda:originna-tive:days_above_springfresh_std	-1.953152	0.468932	-4.17	3.11e-05 ***
wpfgTdr:originna-tive:days_above_springfresh_std	-2.770192	0.472856	-5.86	4.67e-09 ***

```
cover_ar_TMBmod_1 <- glmmTMB::glmmTMB(
    hits ~ log_hits_tm1 +
    days_above_baseflow_std * wpfg * origin +
    days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std^2 +
    # zone * period +
    # zone + period +
    # grazing + wpfg  +
    (1 | site / transect) +
    # (1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
    # offset(npoint),
    family = poisson,
    ziformula = ~ wpfg,
    # dispformula = ~ wpfg ,
    data = veg_cover_ar_sum |>
        filter(!wpfg_ori %in% c("Atl_native",
                                "Ate_native",
                                "Tda_unknown"))
)
```

Listing D.2: Vegetation cover, regime model structure.

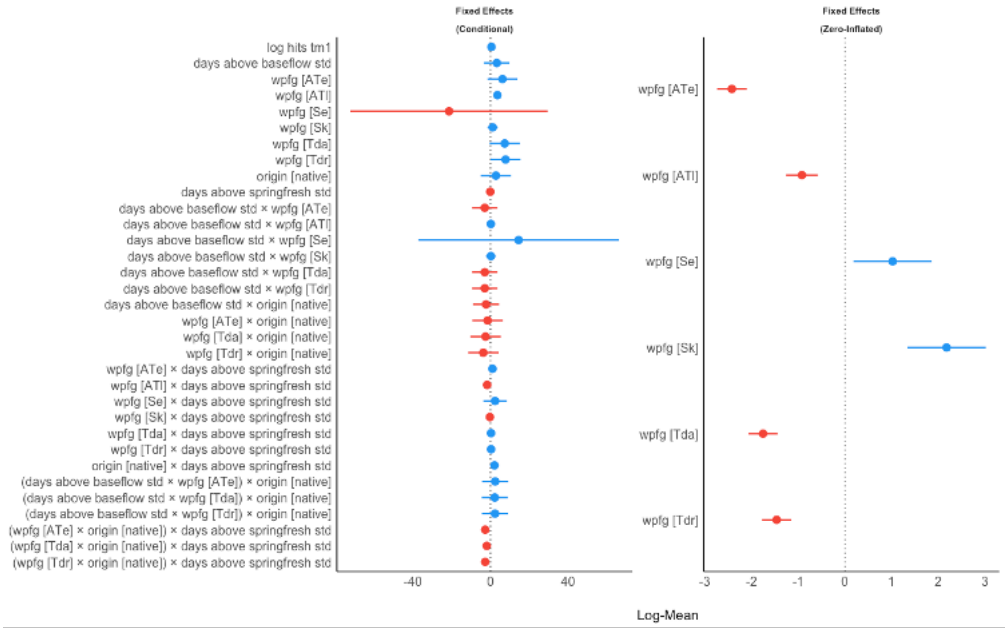


Figure D.1: *Effects plot for vegetation cover flow regime model.*

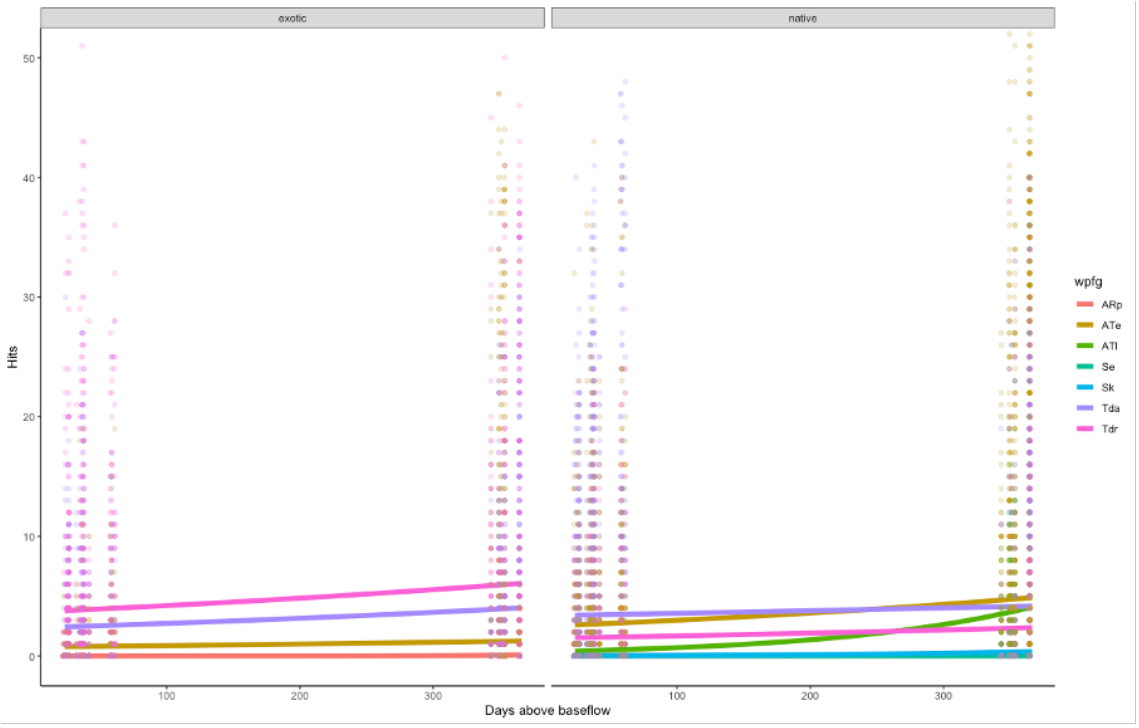


Figure D.2: *Model predicted plant cover for the three way interaction between `Days above baseflow`, `functional group` and `origin`. Points are raw data.*

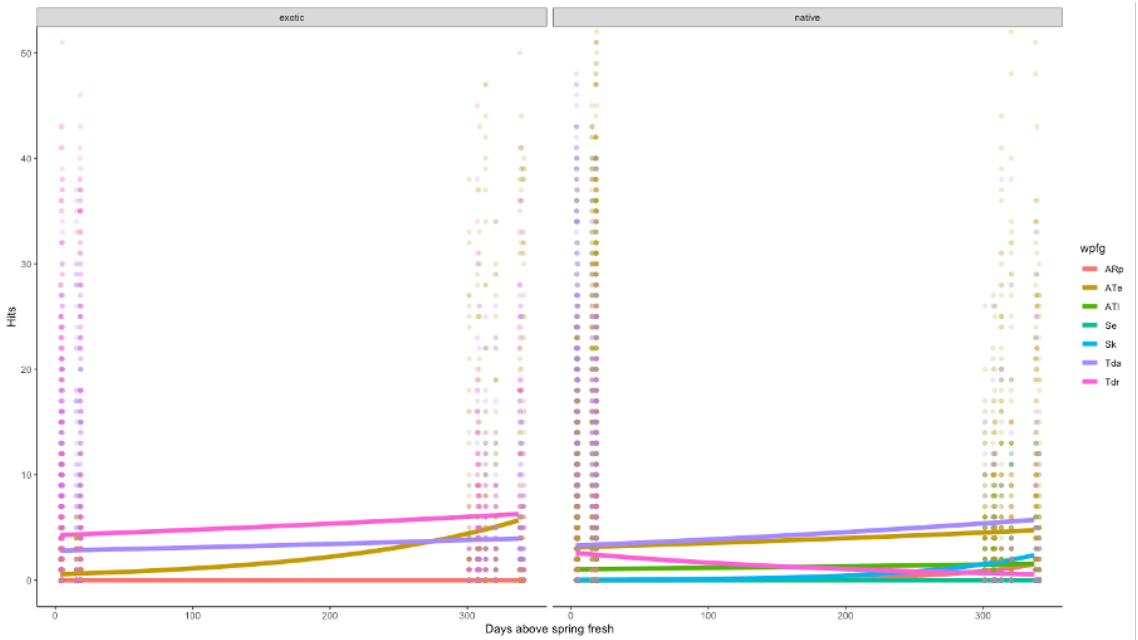


Figure D.3: *Model predicted plant cover for the three way interaction between `Days above springfresh`, `functional group` and `origin`. Points are raw data.*

```
cover_ar_TMBmod_2 <- glmmTMB::glmmTMB(
    hits ~ log_hits_tm1 +
    # days_above_baseflow_std * wpfg * origin +
    # days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std_sq +
    zone * period +
     origin + wpfg +
    grazing +
    (1 | site / transect) +
    # (1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
    # offset(npoint),
    family = poisson,
    ziformula = ~ wpfg,
    # dispformula = ~ wpfg ,
    data = veg_cover_ar_sum |>
    filter(!wpfg_ori %in% c("Atl_native",
                            "Ate_native",
                            "Tda_unknown"))
)
```

Listing D.3: Vegetation cover, flow event model (version 1).

D.6.3 Vegetation Cover: flow event model (version 1)

This model includes several categorical predictors (`origin`, `wpfg`, `grazing`, and a `zone`-by-`period` interaction) and the full suite of random effects. The aim is to examine how vegetation cover changes in specific zones before and after key flow events (spring and summer freshes). This first version allows functional groups to have different levels of cover *but assumes that changes following flow events are similar in all groups within a zone*. Although this is not an ideal model structure, it is likely that functional groups are restricted to particular zones, in which case the `zone`-by-`period` interaction may capture some of the variation attributable to functional groupings.

The model indicated that vegetation cover differed in its responses to spring and summer freshes in each zone (below baseflow, baseflow to spring fresh, above spring fresh). Below baseflow level, vegetation increased following the spring fresh and remained high following the summer event (Figure D.5). In the baseflow-to-spring fresh zone, vegetation increased following the spring fresh but returned to pre-spring levels following the summer fresh (Figure D.5). In the above-spring fresh zone, vegetation level declined following both the spring and the summer fresh (Figure D.5).

Table D.2: *Vegetation cover flow event model (version 1) summary. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
(Intercept)	0.507407	0.202834	2.50	0.0124 *
log_hits_tml	0.503130	0.004658	108.00	< 2e-16 ***
zonebaseflow_to_springfresh	0.135690	0.019849	6.84	8.14e-12 ***
zonebelow_baseflow	0.027219	0.062187	0.44	0.6616
periodafter_summer	-0.095429	0.019586	-4.87	1.10e-06 ***
periodbefore_spring	0.275100	0.019116	14.39	< 2e-16 ***
originnative	0.022715	0.011601	1.96	0.0502 .
wpfgATe	0.568841	0.074505	7.63	2.26e-14 ***
wpfgATl	0.392510	0.077440	5.07	4.01e-07 ***
wpfgSe	-0.532349	0.235144	-2.26	0.0236 *

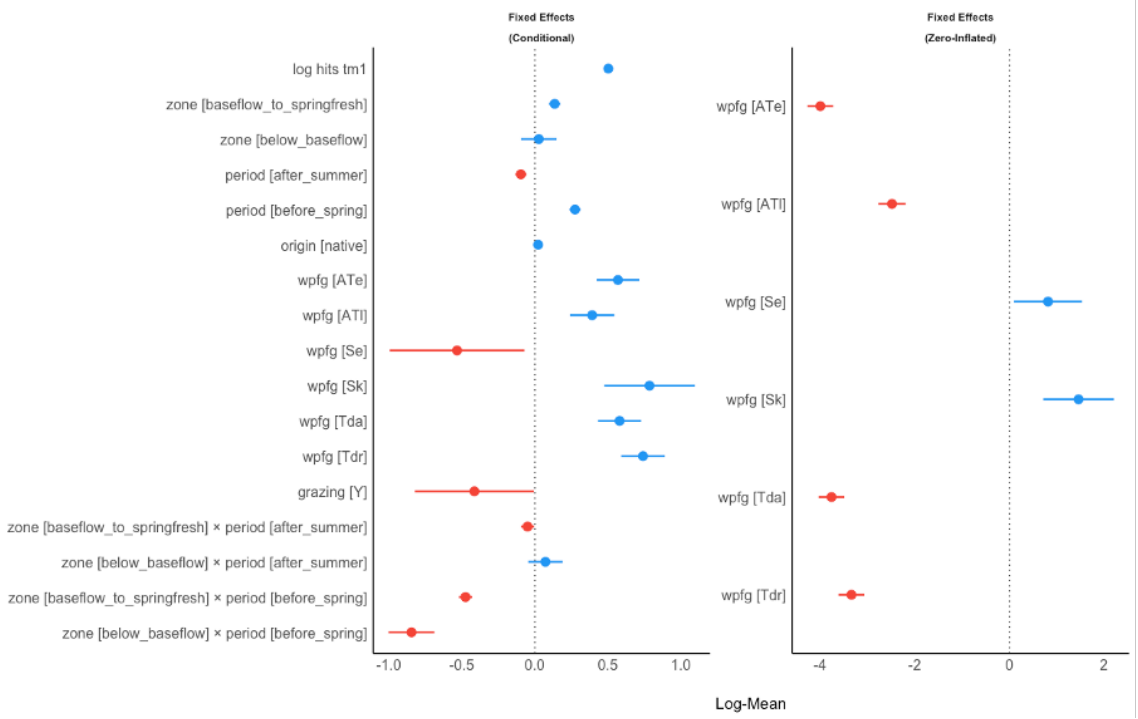


Figure D.4: *Vegetation cover flow event model (version 1) effects plot.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
wpfgSk	0.785098	0.158281	4.96	7.04e-07 ***
wpf gTda	0.580179	0.075459	7.69	1.49e-14 ***
wpfgTdr	0.740681	0.076091	9.73	< 2e-16 ***
grazingY	-0.414008	0.207655	-1.99	0.0462 *
zonebaseflow_to_springfresh:periodafter_summer	-0.049931	0.022829	-2.19	0.0287 *
zonebelow_baseflow:periodafter_summer	0.072623	0.060393	1.20	0.2292
zonebaseflow_to_springfresh:periodbefore_spring	-0.474324	0.023176	-20.47	< 2e-16 ***
zonebelow_baseflow:periodbefore_spring	-0.844370	0.080058	-10.55	< 2e-16 ***

D.6.4 Vegetation Cover: flow event model (version 2)

This model includes several categorical predictors (`origin`, `zone`, `grazing`, and a `wpfg`-by-`period` interaction) and the full suite of random effects. The aim is to examine how vegetation cover of each functional grouping changes before and after key flow events (spring and summer freshes). This second version allows functional groups to have different responses to flow events (spring and summer freshes) *but assumes that vegetation in all zones changes similarly following each flow event*. As for version 1, this is not an ideal model structure but provides an avenue to tease apart `wpfg`-specific responses to flow events.

The model indicated that cover of each functional group differed in its response to spring and summer freshes (Figure D.7). Key patterns were as follows:

- `ARp`: increased after the spring fresh and again after the summer fresh.

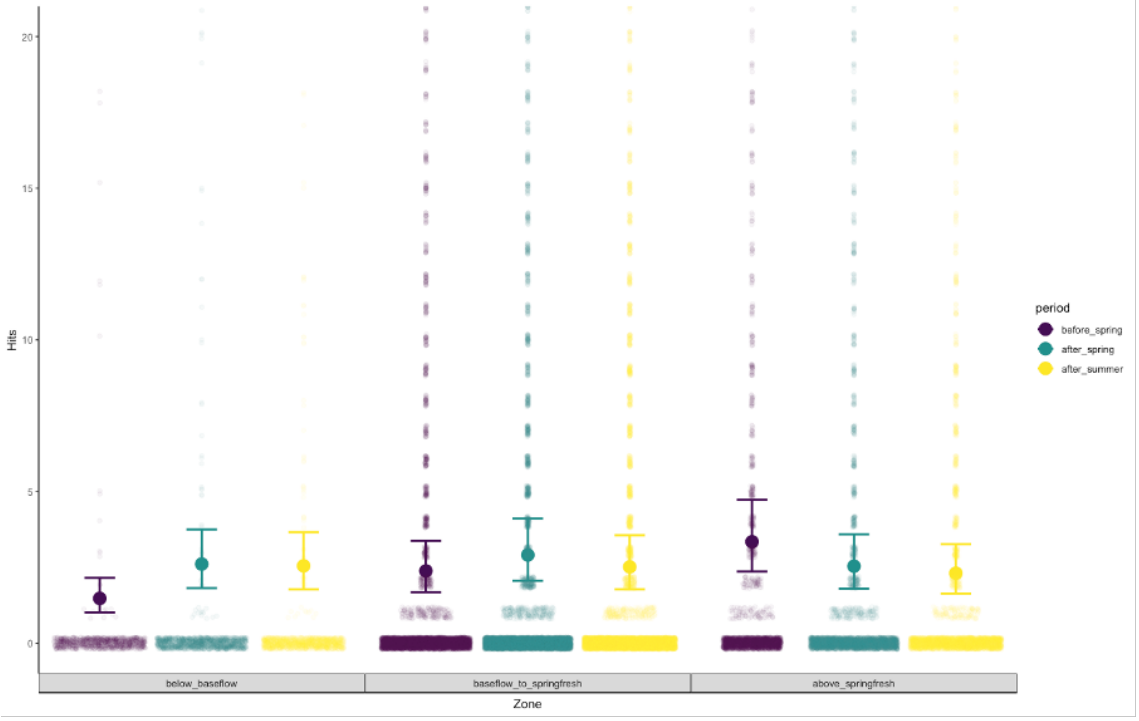


Figure D.5: *Model predicted plant cover for the two way interaction between `period` and `zone`. Points are raw data with the spread across x-axis representing data density.*

```
cover_ar_TMBmod_3 <- glmmTMB::glmmTMB(
    hits ~ log_hits_tm1 +
    # days_above_baseflow_std * wpfg * origin +
    # days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 +
    # days_above_springfresh_std_sq +
    zone + wpfg * period +
    origin +
    grazing +
    (1 | site / transect) +
    #(1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
    # offset(npoint),
    family = poisson,
    ziformula = ~ wpfg,
    #dispformula = ~ wpfg ,
    data = veg_cover_ar_sum |>
        filter(!wpfg_ori %in% c("Atl_native",
                                "Ate_native",
                                "Tda_unknown"))
)
```

Listing D.4: Vegetation cover, flow event model (version 2).

- `ATe`: increased after the spring fresh and returned to pre-spring levels after the summer fresh.
- `ATl`: as for `ATe`.
- `Se`: increased after the spring fresh and largely maintained this level after the summer fresh.
- `Sk`: no change following the spring fresh but large increases following the summer fresh.
- `Tda`: no change following either flow event.
- `Tdr`: reduced following spring fresh then recovered slightly following the summer fresh, but not to pre-spring levels.

Table D.3: *Vegetation cover, flow event model (version 2) summary. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
(Intercept)	0.490933	0.220532	2.23	0.026006 *
log_hits_tml	0.526108	0.004773	110.22	< 2e-16 ***
zonebaseflow_to_springfresh	-0.020485	0.014807	-1.38	0.166522
zonebelow_baseflow	-0.135442	0.051703	-2.62	0.008803 **
wpfgATe	0.762082	0.120587	6.32	2.62e-10 ***
wpfgATL	0.826599	0.124319	6.65	2.95e-11 ***
wpfgSe	-0.219954	0.361766	-0.61	0.543187
wpfgSk	-0.409777	0.497890	-0.82	0.410493
wpfgTda	0.640675	0.121310	5.28	1.28e-07 ***
wpfgTdr	0.361223	0.122625	2.95	0.003222 **
periodafter_summer	0.259131	0.152882	1.69	0.090080 .
periodbefore_spring	-0.769916	0.262410	-2.93	0.003346 **
originnative	0.022738	0.011705	1.94	0.052058 .
grazingY	-0.406325	0.203425	-2.00	0.045780 *
wpfgATe:periodafter_summer	-0.570432	0.153550	-3.71	0.000203 ***
wpfgATl:periodafter_summer	-0.864219	0.159835	-5.41	6.41e-08 ***
wpfgSe:periodafter_summer	-0.445005	0.462139	-0.96	0.335586
wpfgSk:periodafter_summer	1.502212	0.526754	2.85	0.004347 **
wpfgTda:periodafter_summer	-0.310229	0.153846	-2.02	0.043749
wpfgTdr: periodafter_summer	0.052045	0.155180	0.34	0.737333
wpfgATe:periodbefore_spring	0.414490	0.262945	1.58	0.114948
wpfgATl:periodbefore_spring	-0.060499	0.268355	-0.23	0.821633
wpfgSe:periodbefore_spring	-0.968411	0.865331	-1.12	0.263088
wpfgSk:periodbefore_spring	0.760442	0.736038	1.03	0.301531
wpfgTda:periodbefore_spring	0.611485	0.263133	2.32	0.020133 *
wpfgTdr:periodbefore_spring	1.539382	0.263473	5.84	5.14e-09 ***

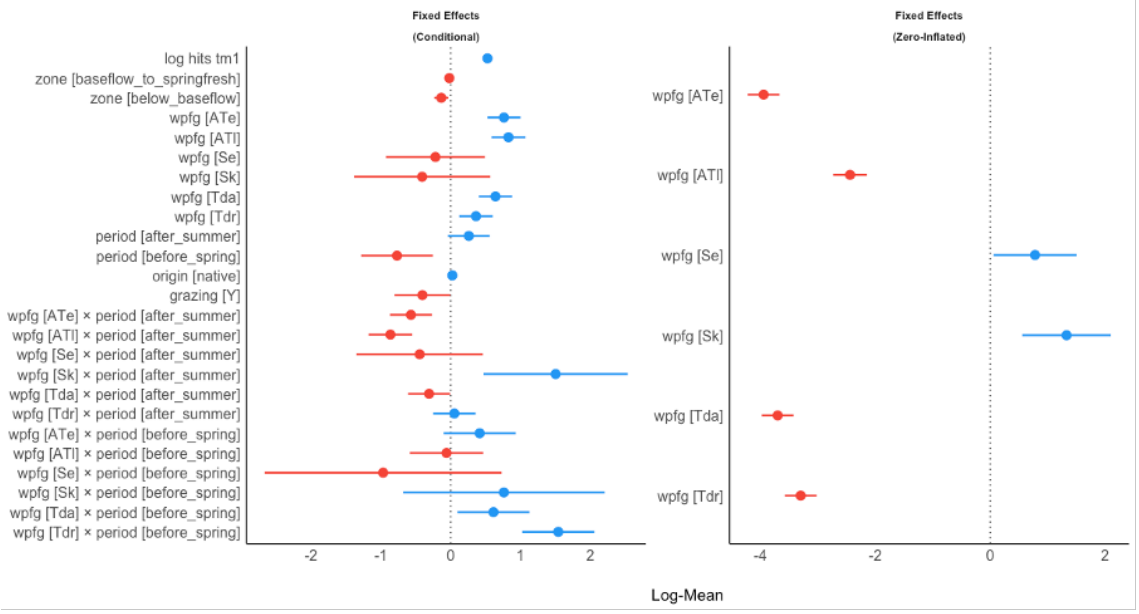


Figure D.6: *Vegetation cover, flow event model (version 2) effects plot.*

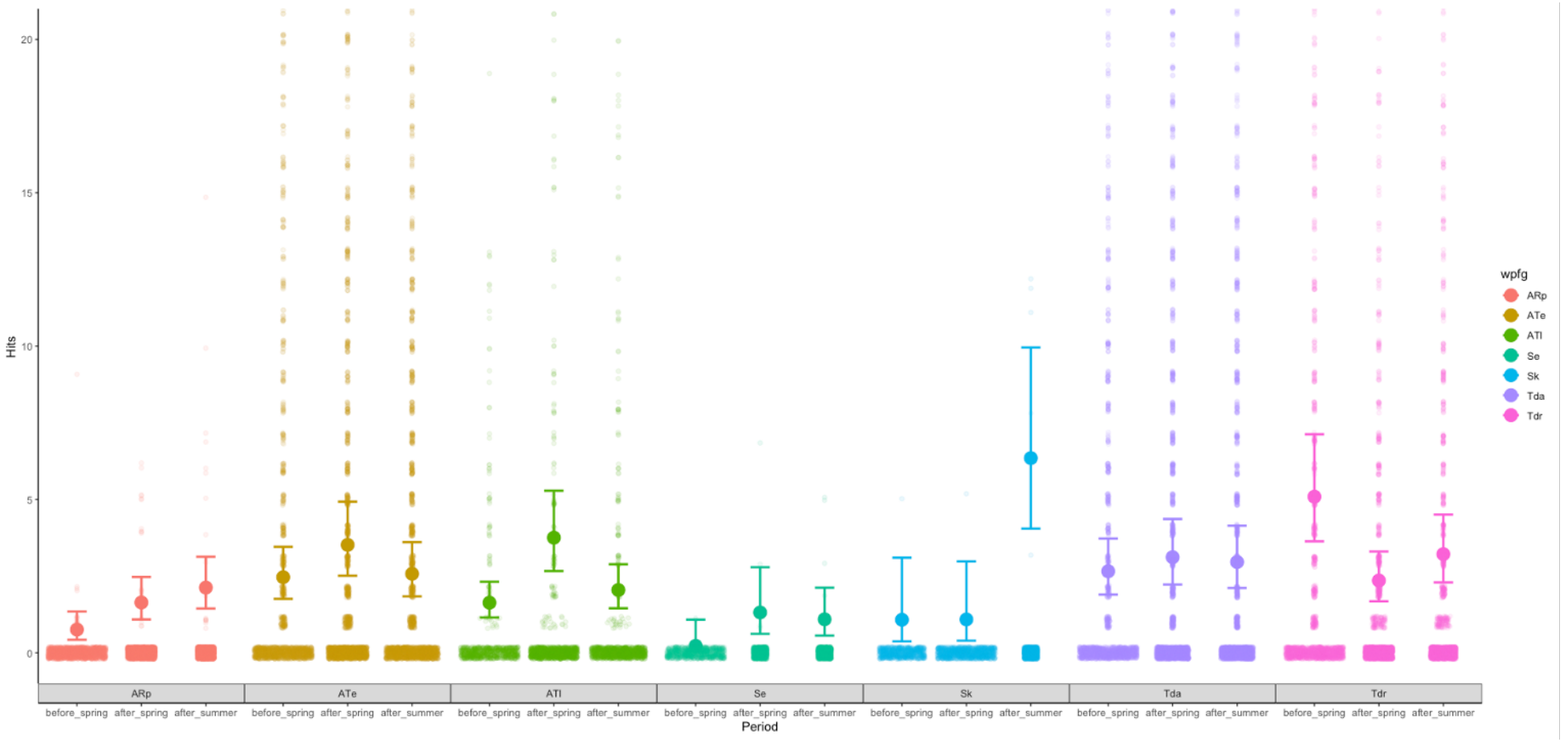


Figure D.7: *Model predicted plant cover for the two way interaction between* `period` *and* `functional group`*. Points are raw data with the spread across x-axis representing data density.*

```
richness_ar_TMBmod_1 <- glmmTMB::glmmTMB(
    richness ~
    days_above_baseflow_std * wpfg * origin +
    days_above_springfresh_std * wpfg * origin +
    # days_above_baseflow_std^2 + days_above_springfresh_std^2 +
    # zone * period +
    # zone * period + zone * wpfg + wpfg * period +
    # grazing + origin +
    (1 | site / transect) +
    # (1 | site / period) +
    (1 | metres) +
    (1 | survey_year),
    # offset(npoint),
    family = poisson,
    # family = nbinom2,
    # ziformula = ~ wpfg,
    # dispformula = ~ wpfg,
    data = veg_richness |>
        filter(!wpfg_ori %in% c("Atl_native",
                                "Ate_native",
                                "Tda_unknown"))
)
```

Listing D.5: Species richness model structure.

D.6.5 Species Richness Model

Species richness models are not fully determined yet. The below structure captures similar patterns to version 1 and version 2 of the cover model. The model includes two-way interactions between `zone` and `period`, `zone` and `plant functional group`, and `period` and `plant functional group`. Together, these terms aim to identify responses to flow events within a given zone or functional group, while accounting for differences in species richness of each functional group among zones.

The models illustrate a general pattern of relatively low temporal variation in species richness once accounting for spatial patterns due to zone and functional grouping. Figure D.8 illustrates that total species richness does not change substantially within each zone following the spring or summer fresh. Similarly, Figure D.10 shows that total species richness of each plant functional group does not change substantially following spring or summer freshes. By contrast, Figure D.10 illustrates that the species richness of each plant functional group differs among zones. As expected, aquatic groupings are more species-richness below the baseflow level, whereas terrestrial groupings are more species richness above the spring fresh level.

JDY comment on Listing D.5: HW to update, the `dispformula` is probably ignored for a Poisson model, use `ziformula` instead. Can flow metrics be included?

Best fitting so far – the three-way interaction had convergence issues.

Table D.4: *Species richness model summary. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
(Intercept)	0.490933	0.220532	2.23	0.026006 *
log_hits_tml	0.526108	0.004773	110.22	< 2e-16 ***
zonebaseflow_to_springfresh -0.020485		0.014807	-1.38	0.166522
zonebelow_baseflow	-0.135442	0.051703	-2.62	0.008803 **
wpfgATe	0.762082	0.120587	6.32	2.62e-10 ***
wpfgATL	0.826599	0.124319	6.65	2.95e-11
wpfgSe	-0.219954	0.361766	-0.61	0.543187
wpfgSk	-0.409777	0.497890	-0.82	0.410493
wpfgTda	0.640675	0.121310	5.28	1.28e-07 ***

Conditional model	Estimate	z value	Std. Error	Pr(>|z|)
wpfgTdr	0.361223	0.122625	2.95	0.003222 **
periodafter_summer	0.259131	0.152882	1.69	0.090080
periodbefore_spring	-0.769916	0.262410	-2.93	0.003346 **
originnative	0.022738	0.011705	1.94	0.052058
grazingY	-0.406325	0.203425	-2.00	0.045780 *
wpgATe: periodafter_summer	-0.570432	0.153550	-3.71	0.000203 ***
wpfgATl:periodafter_summer	-0.864219	0.159835	-5.41	6.41e-08 *
wpfgSe:periodafter_summer	-0.445005	0.462139	-0.96	0.335586
wpfgSk:periodafter_summer	1.502212	0.526754	2.85	0.004347 **
wpfgTda:periodafter_summer	-0.310229	0.153846	-2.02	0.043749 *
wpfgTdr:periodafter_summer	0.052045	0.155180	0.34	0.737333
wpfgATe:periodbefore_spring	0.414490	0.262945	1.58	0.114948
wpfgATl:periodbefore_spring	-0.060499	0.268355	-0.23	0.821633
wpfgSe:periodbefore_spring	-0.968411	0.865331	-1.12	0.263088
wpfgSk:periodbefore_spring	0.760442	0.736038	1.03	0.301531
wpfgTda:periodbefore_spring	0.611485	0.263133	2.32	0.020133 *
wpfgTdr:periodbefore_spring	1.539382	0.263473	5.84	5.14e-09 ***

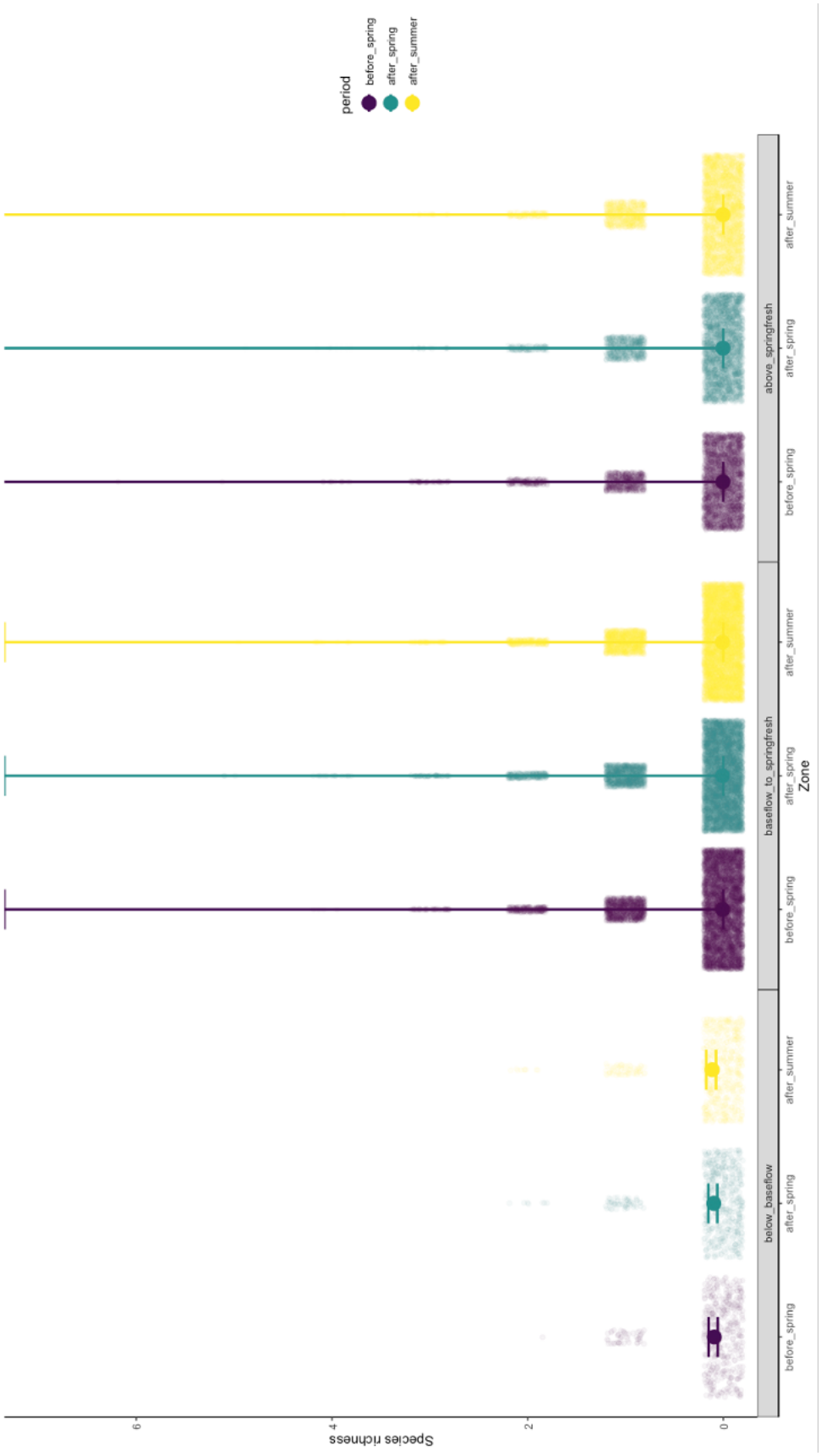


Figure D.8: *Model predicted species richness for the two way interaction between* `period` *and* `zone`*. Points are raw data with the spread across x-axis representing data density.*

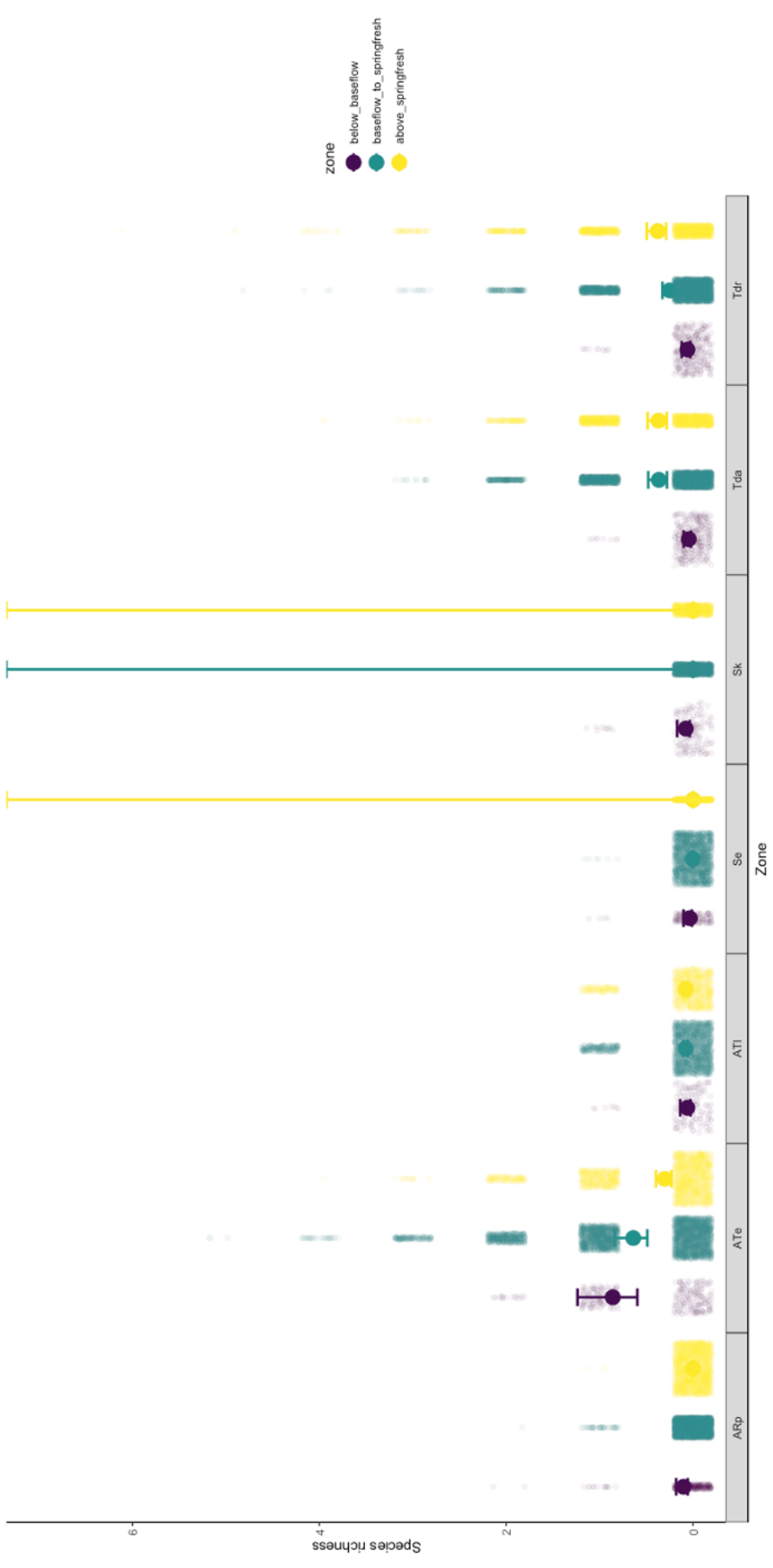


Figure D.9: *Model predicted species richness for the two way interaction between* `zone` *and* `functional group`*. Points are raw data with the spread across x-axis representing data density.*



Figure D.10: *Model predicted species richness for the two way interaction between* `period` *and* `functional group`*. Points are raw data with the spread across x-axis representing data density.*

D.7 Case Study Preregistration with git and GitHub

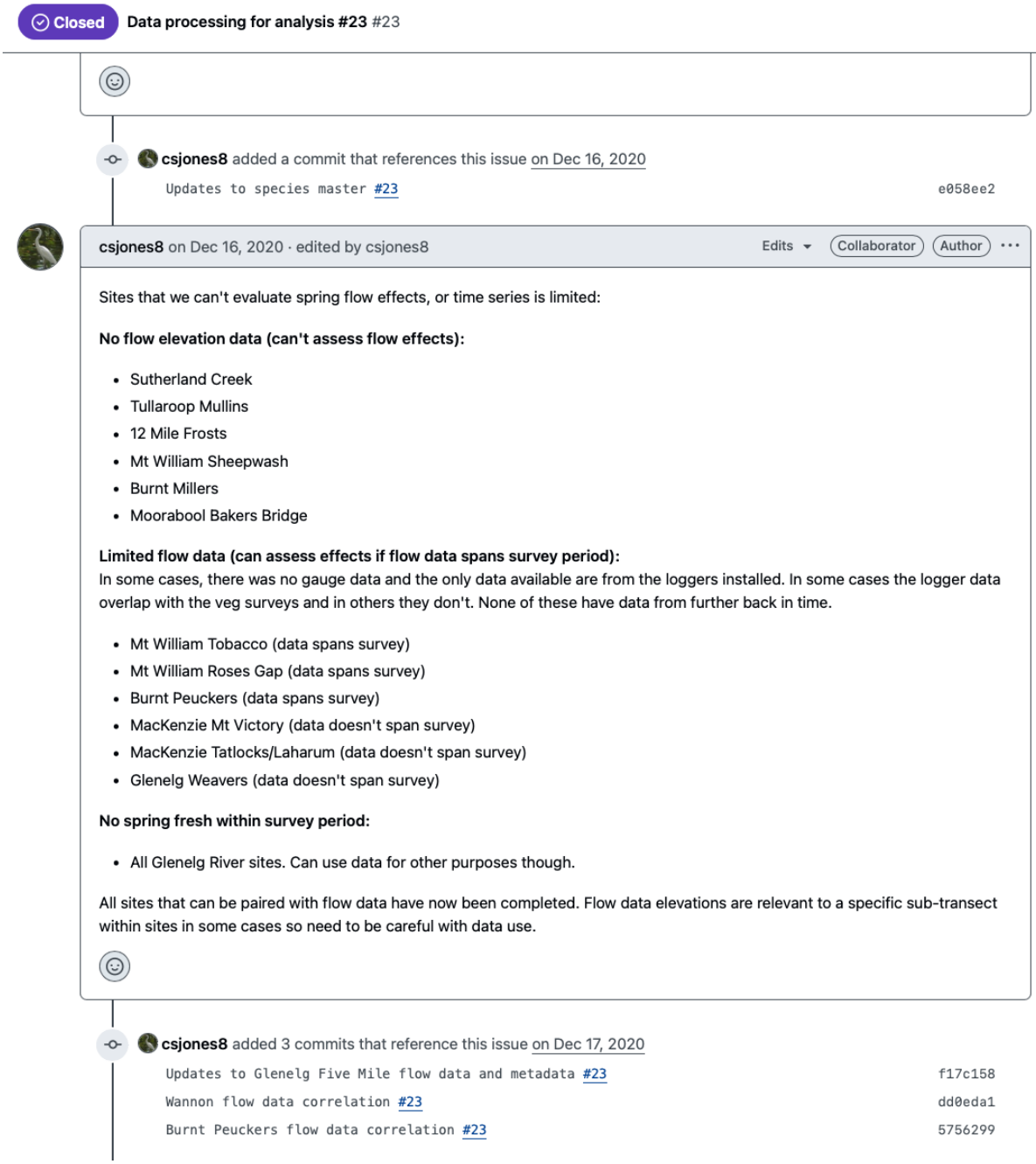


Figure D.11: *Screenshot of GitHub issue thread illustrating where the preregistered analysis tasks and rationale were transparently recorded. The full thread can be viewed at https://github.com/egouldo/VEFMAP_VEG_Stage6/issues/23.*



Figure D.12: *Screenshot of comment on GitHub issue illustrating data-dependent decision-making and where deviation from preregistration occurred during the case study. The full thread can be viewed at https://github.com/egouldo/VEFMAP_VEG_Stage6/issues/71.*

E

Published Article for the "Adaptive Preregistration" Study

The following pages include the published article (Gould, Jones, Yen, Fraser, Wootton, Good, et al., 2026) presented in Chapter 5, which has been reproduced in entirety with permission under license CC BY-NC 4.0:

> **Gould, E.**, Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H. F., Good, M. K., Duncan, D. H., Hauser, C. E., Wintle, B. C., & Rumpff, L. (2026). 'But I can't preregister my research': Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research. Methods in Ecology and Evolution, 2041-210x.70311. https://doi.org/10.1111/2041-210x.70311.

Received: 26 August 2025 | Accepted: 16 March 2026

DOI: 10.1111/2041-210x.70311

Methods in Ecology and Evolution

RESEARCH ARTICLE

'But I can't preregister my research': Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research

Elliot Gould[1] | **Christopher S. Jones**[1,2] | **Jian D. L. Yen**[2] | **Hannah S. Fraser**[1] | **Henry F. Wootton**[2] | **Megan K. Good**[1] | **David H. Duncan**[1] | **Cindy E. Hauser**[2] | **Bonnie C. Wintle**[1] | **Libby Rumpff**[1]

[1]School of Agriculture, Food and Ecosystem Sciences, The University of Melbourne, Parkville, Victoria, Australia

[2]Department of Energy, Environment and Climate Action, Arthur Rylah Institute for Environmental Research, Heidelberg, Victoria, Australia

Correspondence
Elliot Gould
Email: elliot.gould@unimelb.edu.au

Funding information
Department of Education, Australian Government; Department of Energy, Environment and Climate Action of Victoria

Handling Editor: Pen-Yuan Hsing

Abstract

1. Preregistration is an open science practice which aims to improve research transparency and mitigate questionable research practices, like cherry-picking results. It helps protect against cognitive biases, like hindsight bias, that can influence how study outcomes are interpreted. There has been little uptake of preregistration in ecology and conservation, arguably because existing preregistration templates focus on null-hypothesis significance testing, whereas ecology and conservation often rely on different types of statistical modelling.
2. We argue that preregistration in model-based research in ecology and conservation is both possible and beneficial, using templates adapted for domain-specific methodologies. We applied a user-centred design approach to translate the concept of preregistration into model-based research practice for ecology and conservation.
3. To better align the internal logic of preregistration with the iterative and non-linear process of ecological modelling, we propose, test and evaluate a methodology for 'Adaptive Preregistration', using a case study of modelling managed water releases ('environmental flows modelling') in regulated rivers for maintaining riparian vegetation condition in Victoria, Australia.
4. This research provides a template and methodology for implementing Adaptive Preregistration of ecological models. Although we focus on ecology and conservation in this paper, the concept of Adaptive Preregistration, and the templates developed here, could be applied to model-based research in other scientific disciplines. Modellers in ecology and conservation need no longer cry 'but I can't preregister my research'.

KEYWORDS
ecological modelling, environmental flows modelling, good modelling practice, metascience, open science, preregistration, reproducibility, transparency

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2026 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

1 | INTRODUCTION

Over the past decade, a 'replication crisis' in some scientific fields has exposed problems with the reliability of research findings (Camerer et al., 2018; Neoh et al., 2023). One such problem is the prevalence of questionable research practices (QRPs), including p-hacking, Hypothesising After Results are Known (HARKing) and cherry-picking (Wicherts et al., 2016). QRPs increase the probability of false positive results, artificially inflate effect size estimates and ultimately reduce the replicability of published findings, while encouraging overconfidence in the precision of results and exacerbating the risk of accepting and propagating false facts (Hoffmann et al., 2021). Researchers' self-reports show that QRPs are surprisingly common across disciplines (Agnoli et al., 2017; Liu et al., 2020; Makel et al., 2021), including in ecology (Fraser et al., 2018; Kimmel et al., 2023). Although self-report surveys of QRPs are heavily focused on statistical significance testing research employing *p*-values, some practices, like selective reporting/cherry-picking, occur and operate analogously in other inferential frameworks.

QRPs are precipitated when researchers opportunistically misuse 'researcher degrees of freedom' without disclosure (Bakker et al., 2018; Wicherts et al., 2016). Researcher degrees of freedom refer to the many decisions and alternative analytic choices researchers make throughout the research process (Wicherts et al., 2016). Analytic decisions are typically subjective but methodologically defensible; however, choices and omissions taken at each 'fork' may collectively and significantly influence the final 'analysis strategy' (sensu Hoffmann et al., 2021) or 'modelling path' (sensu Hämäläinen & Lahtinen, 2016). The set of all (plausible or reasonable) possible decisions that a researcher could make over the entire analysis from beginning to end has been termed collectively as the 'garden of forking paths' (Gelman & Loken, 2013). 'Many-analyst studies', where multiple researchers independently investigate the same research question by analysing the same dataset, have now demonstrated the consequences of the 'garden of forking paths', whereby alternative analytical decisions can lead researchers to make vastly different interpretations of results, with conflicting conclusions (Gould et al., 2025; Silberzahn et al., 2018). QRPs emerge when researchers opportunistically navigate this garden through systematic selection bias, tending to follow and report on only the analytical paths that yield the cleanest, strongest or most favourable outcomes, while leaving alternative paths unexplored or undisclosed.

Recent frameworks for understanding QRPs in ecological modelling show that these practices involve artificially inflating the apparent credibility, accuracy or performance of models through this opportunistic path selection (Gould, Fraser, Wintle, et al., 2026). Modellers may engage in 'model fishing' by trialling multiple model structures (e.g. linear, polynomial, mechanistic) and reporting only those with the best fit statistics; 'S-hacking' by testing various performance metrics (e.g. R^2, AIC, accuracy thresholds) and selectively reporting only favourable results; or 'sample curation' by removing inconvenient sites or time periods that do not support expected patterns (Gould, Fraser, Wintle, et al., 2026). When this selective path-taking is repeated across studies, it contributes to publication bias, where negative or inconclusive results remain unpublished in the file drawer, culminating in a distorted literature that masks the true uncertainty and variability in ecological systems while making published models appear more credible and reliable than they actually are.

Good intentions and simply being aware of the potential for researcher degrees of freedom to lead to QRPs cannot sufficiently mitigate their risk because of the unconscious nature of their origin (Zvereva & Kozlov, 2021). Preregistration is an open science practice that aims to distinguish between planned and ad hoc analyses (Parker et al., 2019), thereby restricting opportunities for trialling analytical alternatives ('researcher degrees of freedom') and making any deviations from the preregistered plan transparent (Wicherts et al., 2016). Preregistration requires researchers to register their methods and analysis plan in a secure and publicly accessible platform prior to collecting and/or analysing data, which cannot be altered after submission (Parker et al., 2019). Preliminary empirical evidence shows that studies with preregistered analysis plans are more transparently reported and are less likely to contain positive or significant results (Brodeur et al., 2024). This finding in part reflects the researcher commitment to honest, complete and transparent reporting, especially when the preregistration template includes detailed instructions and requirements (Bakker et al., 2018), and a correction for the selective reporting practices that contribute to publication bias. Preregistering studies can also combat the impact of publication bias by making studies findable regardless of whether they are published, in a similar way to a medical trial registry. 'Registered Reports', a publication format that integrates preregistration into the peer-review process (Koivisto & Mäntylä, 2024), contain a direct mechanism for addressing publication bias, whereby the methods and analysis plan are peer-reviewed in advance and the decision to publish is made *before* the results are known.

One potential objection to preregistration for modelling is that model development is an intrinsically exploratory process (MacEachern & Van Zandt, 2019), while preregistration is for confirmatory research (i.e. hypothesis or theory testing, Prosperi et al., 2019). We acknowledge that preregistration serves a crucial role in documenting research intentions and preventing HARKing and other QRPs, and that the exploratory/confirmatory distinction provides valuable guidance where it applies clearly. For model-based research that can be fully specified in advance—where analytical strategies are well-established and data-dependent decisions are minimal—'static' preregistration remains the preferred approach. These conceptions of modelling and preregistration are not entirely wrong, but they are oversimplified for ecological modelling. In practice, there is not such a clear distinction between exploratory and confirmatory activities (Alspaugh et al., 2019; Devezer et al., 2021; Flournoy et al., 2020). Modelling includes some tasks and activities that are exploratory in nature but can be distinguished from wholly exploratory analysis in that modelling is often goal-driven and

directed (Alspaugh et al., 2019). Researchers move back and forth between different types of reasoning and inference throughout the modelling process, with deductions leading to abductions as the analysis proceeds (Alspaugh et al., 2019; Prosperi et al., 2019) or even occurring together in a hybridised manner (Prosperi et al., 2019).

In this iterative and non-linear process of modelling, initial analyses often inform future analyses of the same data, whereby the modeller's understanding of the problem and the model itself are incrementally adjusted in light of intermediate results from previous decision points (Dwork et al., 2015; Hämäläinen & Lahtinen, 2016). Recent evidence from many-analyst studies supports this view: Schweinsberg et al. (2021) found that even within individual analyses, researchers engage in both exploratory and confirmatory activities iteratively, contrasting against definitions of 'confirmatory' analysis used within the preregistration literature, where confirmatory analyses are typically 'planned out and 'frozen' online prior to having the dataset'. Modellers are typically uncertain about which model is most appropriate for the data (Popovic et al., 2024) and are reluctant to register a modelling strategy that may turn out to be incompatible with the data (Roettger, 2019). For example, modellers may perform exploratory analyses to validate distributional assumptions (Campbell, 2021), which may result in subsequent changes to the model specification. We argue that systematic data-dependent decisions like these should not be interpreted as violating a key tenet of preregistration, that is, failing to define the entire analysis protocol ahead of time (Dwork et al., 2015; MacEachern & Van Zandt, 2019). The key distinction is not whether decisions are data-dependent, but whether they are systematic (following preregistered decision rules) versus opportunistic (result-seeking without disclosure, see Gould, Fraser, Wintle, et al., 2026).

We propose a methodology for implementing an expanded view of preregistration called 'Adaptive Preregistration' (sensu Srivastava, 2018), which captures decision points and reasoning about modelling choices, and embraces the data-dependency inherent within the modelling development process for research contexts that don't fit neatly into binary exploratory/confirmatory classifications. This goes beyond the work of others who offer discussion and guidance about when and how to deviate from a preregistration (Lakens, 2024; Willroth & Atherton, 2024), instead providing an approach to preregistration embodying the principles of 'registered flexibility' (Roettger, 2019; Srivastava, 2018) and iterative or 'interim preregistrations' (Hofman et al., 2023; Ioannidis, 2022).

In this study, we aim to translate preregistration for application to model-based research, specifically aimed at ecology, conservation and related disciplines. Here, we first identify a generalised 'modelling workflow' that captures critical analysis decisions in the model development cycle and use it to build an Adaptive Preregistration template that can work for ecological modelling. We then test and evaluate our template using a case study of modelling managed water releases in regulated rivers ('environmental flows') to improve riparian vegetation condition in Victoria, Australia. This case study demonstrates a novel approach to preregistration using a template designed for model development and evaluation.

2 | MATERIALS AND METHODS

We modified preregistration for use in model-based research contexts following a user-centred design framework (Pavelin et al., 2012), where users were engaged at all stages to produce a preregistration template and protocol that is fit for purpose (Figure 1). All study materials are archived at osf.io/sq5j4/ (Gould, Jones, et al., 2026).

2.1 | Problem formulation: Defining vision and objectives for a preregistration template for ecological modelling

We first specified the research context, purpose, target users and use-cases (see Figure 1, step 1; Pavelin et al., 2012; Pu et al., 2019). Our focus is model-based research, which we define broadly as any research that uses quantitative modelling to answer its research question (Brudvig, 2017). This encompasses diverse modelling goals (exploration, inference, and both explanatory and anticipatory prediction; Bodner et al., 2020; Mouquet et al., 2015; Tredennick et al., 2021), varied research contributions (developing new methods, applying or extending existing approaches, and model transfers; Yates et al., 2018) and different model types (phenomenological vs. mechanistic; Connolly et al., 2017) across basic and applied research contexts and disciplines (Brudvig, 2017; Lewis et al., 2023). We intentionally maintain this broad definition because ecological modelling is highly diverse in its methodological approaches, analytical purposes and research applications.

The purposes of the preregistration template and methodology are to:

1. Delimit 'researcher degrees of freedom' (Pu et al., 2019) to mitigate the risk of QRPs when conducting model-based research in applied ecology and conservation.
2. Increase 'research transparency' of ecological modelling, including both:
   (i) *Production transparency*—including research artefacts like open-access data and materials, or data collection procedures (Lupia & Elman, 2014).
   (ii) *Analytic transparency*—a complete account of how analytic conclusions are drawn from the data (Lupia & Elman, 2014) where model choices, steps in the modelling process, assumptions and expectations about model outputs are clearly articulated (Bodner et al., 2020).

We specified two use-cases for our Adaptive Preregistration template:

1. Study authors or modellers seeking to preregister their modelling study.
2. Editors, reviewers, and/or readers of the completed preregistration, who are tasked with 'preregistration checking' or verifying that analyses and findings reported in the final manuscript are consistent with the preregistered analysis plan (Pu et al., 2019).

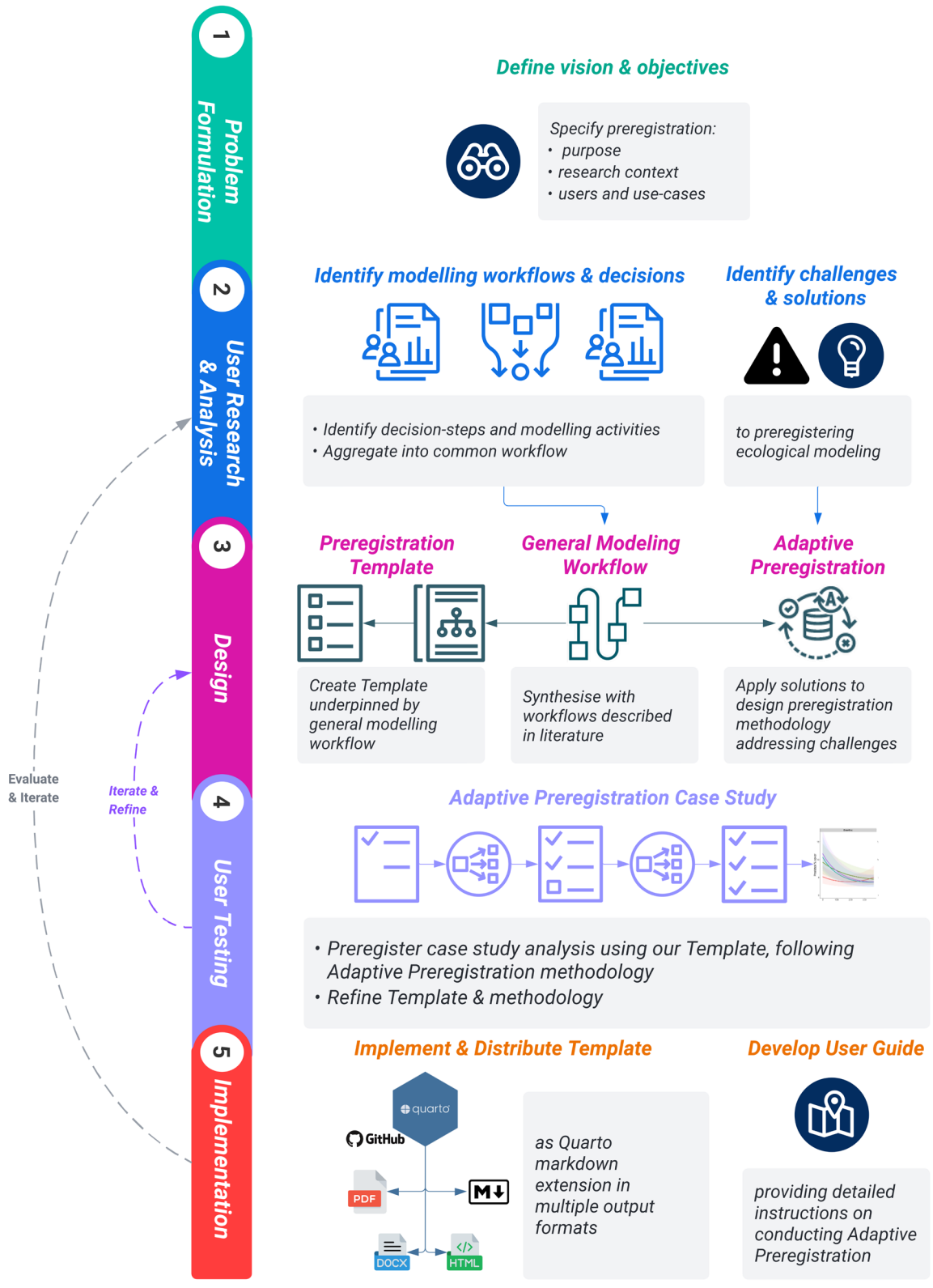


FIGURE 1 Process for translating preregistration to ecological modelling research. We followed the user-centred design steps proposed by Pavelin et al. (2012), except for the 'Evaluate and Iterate' phase of the process, which we hope will be taken up by the community when trialling the practical tools developed in this study. Following a collaborative workshop (Step 2), we specified a general modelling workflow that informed the structure and content of a preregistration template (Step 3) and proposed a methodology we call 'Adaptive Preregistration' (Step 3) addressing challenges to preregistering model-based research identified in Step 2. We refined the template and methodology using a real-world modelling case-study (Step 4). The template is provided in multiple formats (Step 5), including as a Quarto Markdown extension (Gould, 2026), and instructions for conducting Adaptive Preregistration are detailed in a user guide (Gould, Fraser, Jones, et al., 2026; Appendix S3).

Existing preregistration formats are designed primarily for study authors; they are static text-based documents that do not easily facilitate preregistration checking. We designed our template with editors, reviewers and readers in mind.

2.2 | Designing the general modelling workflow

Building on our broad definition of model-based research (Section 2.1), we conducted a literature review (Appendix S1) to identify a modelling workflow suitable for application to a range of different modelling goals, research contributions, model types and problem contexts. The final output was a workflow that reflected a balance between idealised and actual modelling in practice, aligned with the prescriptive guidelines of 'good modelling practice' (Augusiak et al., 2014).

2.3 | Designing the preregistration template

We organised and facilitated a workshop with 10 ecologists and ecological modellers (Appendix S2). A key task was to identify common modelling workflows and critical decision-steps in model development (Figure 1, step 2). 'Critical' decision-steps are points in the modelling process where researchers, or modellers, make analytical decisions that could change the output of the analysis. In this activity, workshop participants first individually reflected on their own scientific process for a recent or memorable research project, listing the steps undertaken throughout the process. Next, participants collaboratively mapped the decision steps in their own personal modelling workflows onto a modelling workflow template prepared earlier (osf.io/fgd23), informed by our general workflow (Figure 1). We then collaboratively reviewed and refined the modelling workflow into a standard set of modelling phases and commonly implemented decision steps that formed the basis of the preregistration template (Figure 1, step 3).

2.4 | Developing an Adaptive Preregistration methodology

We developed a methodology for application of preregistration to model-based research, or Adaptive Preregistration (Figure 1, step 3), comprising two key components: registered flexibility and interim preregistrations. We summarise the methodology and direct readers to our user-guide for more detailed guidance on implementing Adaptive Preregistration in practice (Gould, Jones, et al., 2026; Appendix S3).

2.4.1 | Registered flexibility

Registered flexibility includes preregistration of 'plans to deploy flexible strategies' (Srivastava, 2018), whereby the preregistration author specifies flexible heuristics containing alternate analysis or modelling strategies whose execution depends on the outcome of previous decision-points or analyses. For example, when faced with methodological or analytic uncertainty that cannot be resolved without observing parts of the data, conducting preliminary analyses, model checking or other data-dependent decisions, a modeller can preregister a decision-tree that consists of predefined rules about when a particular modelling strategy or decision should be implemented (Baldwin et al., 2022). There are three requirements for adequately registering flexible analyses:

1. Stating what quantity needs to be known to move forward with the modelling and analysis, and why.
2. Describing the analysis that will be used to generate this quantity and which parts of the data will be used.
3. Explaining how the results will be interpreted, listing each alternative decision under consideration and the analysis result that will trigger each decision, respectively.

2.4.2 | Interim preregistrations

The modeller follows an iterative process of preregistration in parallel with modelling, consisting of interim preregistrations that mark key phases of modelling and analysis as different parts of the data are observed (Srivastava, 2018). As the modeller proceeds through the model development process, they shift from ideation and preregistration to execution of the preregistered analysis plan, and back again, generating interim preregistrations at different points within the model development process, depending on observed outcomes of the flexible heuristics described in the interim preregistration (Figure 2). The realised modelling path executed from the garden of forking modelling paths will depend on the observed outcomes of the registered flexible analyses at each interim preregistration.

2.4.3 | Transparent documentation of Adaptive Preregistration

Ideally, to transparently document the Adaptive Preregistration process and facilitate preregistration checking of flexible modelling and analysis strategies (one intended use-case for our methodology and template, Section 2.1), the results of any interim and final analyses must be explicitly linked back to the preregistered analysis strategy. Researchers can take simple steps to achieve this, for example, by retaining separate versions of initial, interim and final preregistration documents; retaining the implementation and results of registered flexible analyses alongside the preregistration documents while recording their filenames in the relevant preregistration document; and utilising existing platforms with basic version control features, such as the Open Science Framework (OSF, cos.io/products/osf).



FIGURE 2 Implementing Adaptive Preregistration. The preregistration (blue steps) and modelling and analysis processes (orange steps) operate in parallel. Purple tags represent git tags, corresponding to incremental versions of the preregistration. Any changes to the analysis plan receive a GitHub release and major version number increment. Step 3 corresponds to the initial preregistration document, Step 6 corresponds to an interim preregistration and Step 8 corresponds to the final preregistration. The execution of the Final Preregistration occurs in Step 9. Note that the use of git and GitHub are optional (see Section 2.4.4).

2.4.4 | Optional: Preregistration with git and GitHub

The use of version-control for transparently documenting code-based analyses is considered best-practice, however, is not yet widespread in ecology and related disciplines (Braga et al., 2023). We propose using git and GitHub as the vehicle for transparent documentation of Adaptive Preregistration of a modelling study for those researchers already comfortable with these tools (see user guide for details, Gould, Jones, et al., 2026; Appendix S3). The procedure leverages GitHub's tag and release feature (help.github.com/en/github/administering-a-repository/releasing-projects-on-github) in conjunction with Semantic Versioning (semver.org) to track, document and collaborate on changes to the preregistration document and the analysis. Modelling code and results are stored and versioned alongside the preregistration document in the project repository, while GitHub 'issues' (see Braga et al., 2023 for explanation and definition) are used to track discrete analysis tasks described in the preregistration. GitHub issues are particularly useful for registered flexible analyses, facilitating documentation of any interpretation of the results of registered flexible analyses and their influence on the rationale for the subsequent preregistration; the outcome of the registered flexible analysis is recorded within that issue, hyperlinked to the registered heuristic in the preregistration document, alongside interpretation of those results and documented decisions about which decision alternative is to be selected based on those results.

2.5 | Evaluating and refining Adaptive Preregistration with a case study

We preregistered a real-world model-based research problem as a case study for evaluating the preregistration template and proposed Adaptive Preregistration methodology (Figure 1, step 4). The case study analysed the effectiveness of environmental flows on maintaining riparian vegetation condition in Victoria, Australia (Jones et al., 2025; Jones, Vivian, et al., 2026), forming a component of the Victorian Environmental Flows Monitoring and Assessment Program (VEFMAP), a large-scale, long-term monitoring programme delivered by the Arthur Rylah Institute for Environmental Research. Case study data were collected under a research permit for the Victorian Flora and Fauna Guarantee Act, permit number 10010953. The VEFMAP research team consisted of a project lead (CJ), who was responsible for establishing the conceptual framework and research questions that inform the analysis, and a modeller (HW) who undertook much of the coding and exploratory analysis under the direction of a lead modeller (JDY). Modelling experience among the team ranged from moderate to advanced, and the preregistration was collaboratively completed by the case study research team and the study lead (EG). In addition to following the protocol for Adaptive Preregistration using git and GitHub, we also utilised the version-control and collaborative project management features of git and GitHub (Braga et al., 2023) to live–develop the preregistration template and capture the case study researchers' feedback about the template and Adaptive Preregistration methodology. For example, if it emerged in the process of completing the case study preregistration that we had missed an important step in the modelling process or found that the order and structure of the template should change, the suggested change and its justification was recorded in GitHub discussions. We also conducted follow-up semi-structured interviews with the case study researchers to capture detailed reflections.

3 | RESULTS

3.1 | The preregistration template

The final preregistration template, refined through application of the draft template developed in the user-research workshops to the case study, is hosted at github.com/egouldo/EcoConsModPreReg and provided as a Quarto Markdown extension (Gould, 2026) that can be installed and rendered in html, pdf and docx formats (Appendices S4 and S5). Intermediate versions of the template developed through application of the initial template to the case study are available in the case study preregistration document's GitHub commit history (github.com/egouldo/VEFMAP_VEG_Stage6/commits/master/analysis/preregistration_template.Rmd, osf.io/sq5j4/files/aup5f) or in case-study GitHub repository releases (github.com/egouldo/VEFMAP_VEG_Stage6/releases, osf.io/sq5j4/files/jdbr7).

3.2 | Findings from the case study

The application of our preregistration template and Adaptive Preregistration methodology to the environmental flows modelling case study occurred across three distinct stages (Table 1; Figure 3), with two interim preregistrations (V1 and V2) preceding the finalised preregistration (V3, Appendix S6; Jones, Yen, et al., 2026), each corresponding to distinct phases in both the 'Preregistration' and 'Modelling and Analysis' processes. While preparing Version 1, the modellers decided that they needed to conduct analyses to inform specification of appropriate candidate models. They preregistered a 'Preliminary Analysis' (V2) which used a subset of the full dataset (which was not discarded) to preserve some degree of data-dependent decision-making while permitting a degree of exploratory analysis. The Preliminary Analysis involved exploratory analyses to inform model specification and analysis for the 'Main Analysis' on the full dataset (V3, Appendix S6; Jones, Yen, et al., 2026). Several classes of data-dependent decisions (Liu et al., 2020) emerged during our case study, which were managed using a combination of registered flexibility and interim preregistrations:

- *Data-flow dependencies*: where the output of one decision is the input to another. For example, the models parameterised on the full dataset in the case study's Main Analysis were subject to a suite of model checks as outlined in the decision-trees

Preregistration version	Preregistration process	Modelling and analysis process
Version 1 (v0.10–v0.8.1)	Establish problem context, study background and aims, describe data collection and data cleaning process, articulate modelling objectives and specify candidate models	Data cleaning and preparation undertaken concurrently by lead and supporting researchers due to limited resourcing and short timelines for project deliverables. Candidate models addressing different components of the research question were preregistered. Data structural properties and potential consequences for model fitting were identified
Version 2 (v0.9.0–v.0.12.5)	Describe Preliminary Analysis on data subset and specify registered flexibility in the form of decision heuristics	Critical uncertainties in modelling decisions were articulated (e.g. how do we classify best flow regime based on inundation data?). Additionally, data partitioning, exploratory data analysis, refined candidate models and model checking procedure were preregistered
Version 3 (v0.13.0–v.0.20.1)	Describe Main Analysis, including specification of decision tree for model selection process based on results of Preliminary Analysis	Conducted modelling and analysis as outlined in preregistration. Reported deviations and rationale as necessary

TABLE 1 Key stages of the modelling and Adaptive Preregistration process implemented in the environmental flows modelling case study. Version numbers in parentheses are GitHub tags, marking snapshots of the repository for each Preregistration Version (github.com/egouldo/VEFMAP_VEG_Stage6/releases, osf.io/sq5j4/files/jdbr7).

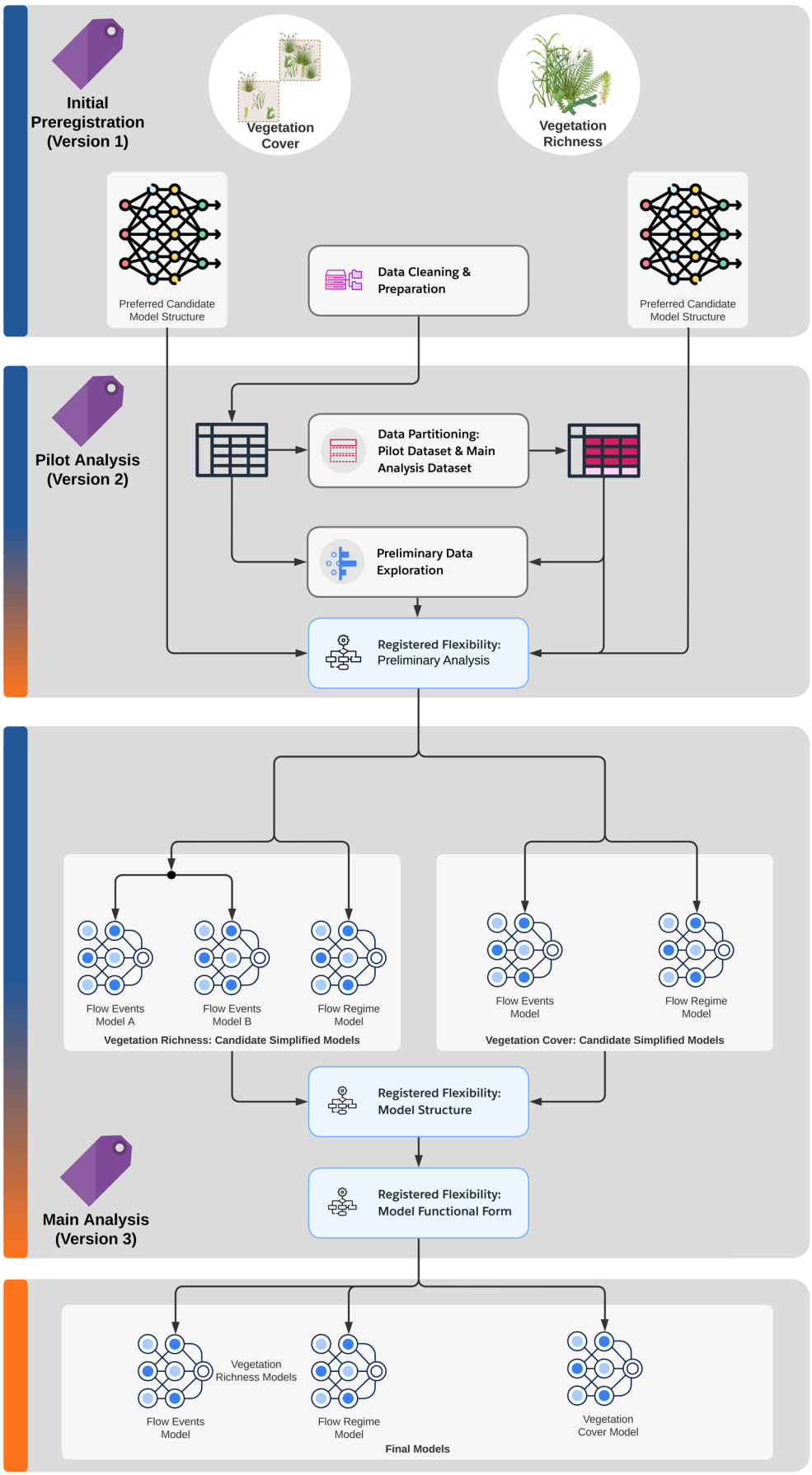
in Preregistration Version 3 (Figure 4), with the results of model checking informing a priori specified alternative model structures and model functional forms.

- *Information dependency*: where one decision informs another. For example, outcomes from the exploratory Preliminary Analysis (Interim Preregistration V2, Table 1) informed the model structures specified in Preregistration Version 3 (Table 1), as well as the decision heuristics for triggering their fitting and acceptance.
- *Procedural dependency*: where downstream decisions would not exist if some alternative upstream decision was made instead. For example, if the Preliminary Analysis had not revealed over-dispersion or zero-inflation in the data, then the case study modellers would not have fitted zero-inflated Poisson models, nor would they have trialled accounting for over-dispersion or refitting using negative binomial models (Figure 4).

3.2.1 | Interim preregistration (Versions 1 and 2)

Two initial candidate 'maximal' models were preregistered in Version 1 (Figure 3), one for each study response variable; however, modellers were unsure if these models would be supported by the dataset due to low sample sizes relative to the complexity and spatio-temporal patchiness of the data. Preregistration Version 2 described an exploratory Preliminary Analysis that aimed to resolve critical uncertainties about how to operationalise the models, given data constraints and how to specify a candidate set of models that could feasibly be fitted to the full dataset (Table 1). This involved partitioning the dataset into a 'Preliminary Analysis' subset for making data-informed modelling decisions, which were then preregistered in a subsequent preregistration of the Main Analysis (V3, Table 1; Figure 3). The Preliminary Analysis dataset consisted of observations from a single site with

FIGURE 3 Evolution of model development across preregistration iterations in the case study. Each preregistration version corresponds to different stages of model development, summarised in Table 1. Maximal models for each response variable were specified in the initial preregistration (V1). Exploratory and preliminary analyses and data partitioning were next preregistered (V2), aiming to resolve uncertainties in model specification given data constraints. The Preliminary Analysis identified three candidate simplified models for vegetation richness and two for vegetation cover. These model structures were preregistered in the final preregistration for the Main Analysis (V3), in conjunction with registered flexibility for determining model structure and functional form. The two registered flexibility steps in the Main Analysis (shown in blue) were informed by data exploration and registered analyses conducted during Preliminary Analysis (V2) and are further described in Figure 4. The Main Analysis was preregistered to be conducted on the full dataset. While conducting the preregistered Main Analysis, three final models were identified for model fitting and analysis, one model for vegetation cover and two vegetation richness models, with one incorporating 'flow events' (spring and summer freshes) and the other incorporating 'flow regime' (days above baseflow and days above spring fresh [per year]).


Initial Preregistration (Version 1)
Vegetation Cover
Vegetation Richness
Preferred Candidate Model Structure
Data Cleaning & Preparation
Preferred Candidate Model Structure
Pilot Analysis (Version 2)
Data Partitioning: Pilot Dataset & Main Analysis Dataset
Preliminary Data Exploration
Registered Flexibility: Preliminary Analysis
Flow Events Model A
Flow Events Model B
Flow Regime Model
Vegetation Richness: Candidate Simplified Models
Flow Events Model
Flow Regime Model
Vegetation Cover: Candidate Simplified Models
Registered Flexibility: Model Structure
Registered Flexibility: Model Functional Form
Main Analysis (Version 3)
Vegetation Richness Models
Flow Events Model
Flow Regime Model
Vegetation Cover Model
Final Models

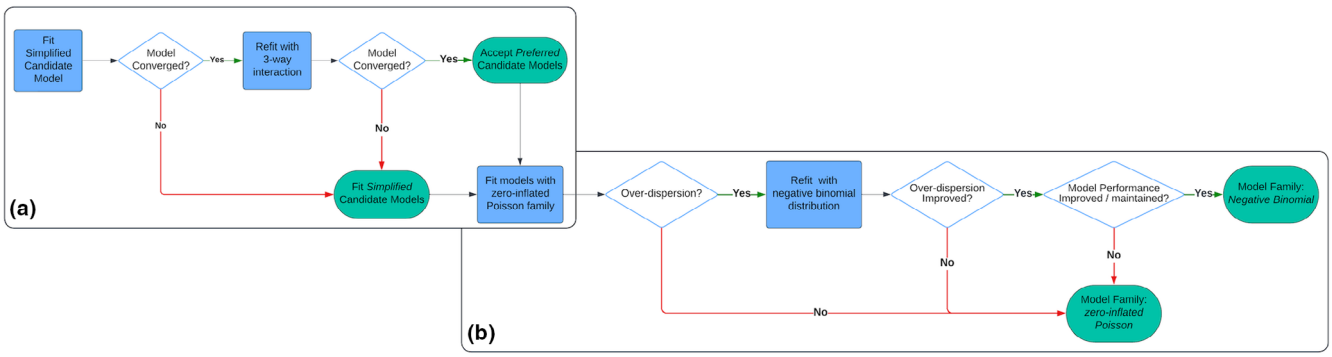


FIGURE 4 Application of registered flexibility within the case study preregistration, for selecting model structure (a) and selecting an alternative functional form over the default zero-inflated Poisson model (b).

good temporal data coverage, while the Main Analysis dataset comprised the full dataset, including the Preliminary Analysis dataset. Preliminary Analysis informed how to operationalise the models given the underlying structure of the dataset. For example, simple tests of data distributions within hierarchies were required to check for data spread and the presence of zero-inflation, while checks for collinearity among candidate predictor variables were needed to prevent confounding and overfitting, and to ensure relevant interactions were captured in the models. The same procedures for evaluating model fit and model checking were preregistered for both the Preliminary and Main Analysis (Figure 4b), however, the purpose of model checking in the Preliminary Analysis was to identify appropriate model structures, whereas the purpose for the Main Analysis was to ensure that the fitted models were supported by the data.

3.2.2 | Final preregistration (Version 3): Main Analysis

In Version 3 of the Preregistration, we derived and preregistered:

1. A preferred candidate model or 'full model' for each response variable (vegetation richness and vegetation cover), to be fitted to the full dataset (Initial Preregistration, Figure 3; Appendix S6, p. 13; Jones, Yen, et al., 2026).
2. A candidate set of simplified models that separately captured different flow components for each response variable, with two versions of the flow events model for vegetation cover (Flow Events Models A & B, Figure 3; Appendix S6, pp. 20–21; Jones, Yen, et al., 2026).
3. Decision-trees for triggering fitting and acceptance of the candidate simplified models should the preferred full model specifications fail to converge and/or provide appropriate parameter estimates (Figure 3; Figure 4a).

   A second flexible strategy, for determining model functional form and to account for potential over-dispersion (Figure 3; Figure 4b). Preregistering this flexibility was motivated by the modellers' concerns that the actual data distributions in the full dataset did not match expectations informed by the Preliminary Analysis. This final decision-point for choosing the model family is subjective and fuzzy, and explicit weightings between the three model performance criteria could not be expressed or preregistered. Instead, modellers explained that the final decision will be guided by the model's overall ability to capture key associations reliably.

3.3 | Adaptive Preregistration in practice: Benefits and challenges

At the end of user testing, case study researchers concluded that preregistration is a valuable tool to facilitate good modelling practice, but also acknowledged several difficulties encountered during Adaptive Preregistration (Table 2).

4 | DISCUSSION

The potential benefits of preregistration are clear: it encourages careful and considered analysis, restricts researcher degrees of freedom and mitigates questionable research practices (QRPs) that can derail research. Preregistration can improve the reproducibility and reliability of scientific research (Koivisto & Mäntylä, 2024).

Existing preregistration templates are best suited to hypothesis testing research, and less suited to the style of analysis often taken by ecologists: that is, iterative modelling of complex relationships, often switching between exploratory and confirmatory practices (Alspaugh et al., 2019; Connolly et al., 2017; Prosperi et al., 2019). Although iteration makes preregistration less approachable for model-based studies, it need not preclude it entirely. Here, we show how Adaptive Preregistration can facilitate best practice via a structured, principled and flexible approach to analysis.

Despite fears that preregistration may limit creativity and flexibility (Pu et al., 2019) and lessen researchers' engagement with data (MacEachern & Van Zandt, 2019), our case study shows how Adaptive Preregistration can improve reproducibility and transparency of process, research outputs and analytic decision-making (Liu et al., 2020). By separating the planning stage from the analysis, our case study researchers found the structured approach focused

TABLE 2 Summary of challenges ⊖, benefits and added value ⊕, as well as tips ▲ for using Adaptive Preregistration for ecological modelling.

Type	Name and description	Case study example	Practitioner recommendations and reflections
▲	**Decide on template resolution** Time required to make preregistration templates specific, precise, and exhaustive may be off-putting. What is an appropriate resolution?	The preregistration template first used in the case study was unwieldy and overwhelming. Resolution was reduced by restructuring, removing and merging template items, making the task more approachable	*'Be willing to modify the template if it does not meet the needs of the project. Any attempt to preregister is better than none at all'*—JDY
▲	**Decide on template formatting** The visual design, presentation and formatting affect usability	Reformatting the template with collapsible help and explanatory text boxes, as well as coloured icons to differentiate preregistration items from explanatory text, improved clarity of the initial preregistration and facilitated more efficient completion of the preregistration using the template	
⊖	**Choose a version-control platform that does not require too much investment in upskilling** Adaptive Preregistration requires interim versions of the preregistration be deposited in a single location that is ideally version controlled. Some platforms have a steeper learning curve than others, impacting on the efficiency of preregistration	Using GitHub to track iterations in the preregistration proved difficult to combine with a GitFlow approach for simultaneously collaborating on the analysis code and preregistration across branches. The lead modeller (JDY) regularly uses git for version control but was less familiar with the GitHub web GUI. Under broad direction of the lead modeller, a second modeller (HW) undertook much of the coding and exploratory analysis, but had limited familiarity with version-control systems, and was unfamiliar with collaborative code-base development tools, such as GitHub. Substantial project time was allocated to developing a collaborative GitHub workflow and familiarising both modellers with this approach	
▲	**Ensure that modelling decisions and registered flexibility are accurately documented and easily discernible when preregistration checking** For complex modelling studies, especially where there are multiple models being preregistered and/or where there are multiple models under consideration within a model selection approach, it is difficult to keep track and accurately document analysis decisions, including flexible decision-heuristics	When checking the preregistration to write the case study report, the lead author (EG) found it difficult to clearly identify registered decision-triggers for choosing alternative model functional forms over preferred forms, due to relevant information being dispersed over multiple subsections within the preregistration (Appendix S6, 4.3.1 Quantitative Model Checking, 3.2.2 Choose Model Family, 3.5 Model Assumptions and Uncertainties, 3.4.2 Estimation Performance Criteria, Jones, Yen, et al., 2026). Moreover, some decisions concerning the model selection process were informed by tacit knowledge of the modelling team and not recorded in the preregistration (Section 4.2)	*'I could not figure out why quadratic associations rather than linear associations were fitted at the functional group level when these were not described in the preregistered full model descriptions and pseudo-code. After significant effort, it became apparent that this decision was described across multiple sub-sections'*—EG. *'In hindsight, we would specify the model selection procedure in both the Preliminary Analysis and Main Analysis more clearly. The process we followed was more akin to "make the model more complex again, and make it look more like the full model"'*—JDY
⊖ ▲	**Adaptive Preregistration may be incompatible with model management workflows in real world applications** Operational procedures may constrain analytic decision-making too early in the research timeline and require modellers to be involved earlier and throughout the research. Operational constraints may be easier to overcome in academic rather than government or industry settings	At the case study institution, similar to many organisations working on applied problems, modellers are typically brought into the research project late in the research programme and are allocated limited time over a discrete period (e.g. 5 days immediately before report-writing and well after project planning) to conduct the modelling with limited knowledge of data provenance, dependency structure and underlying distributions. Following the Adaptive Preregistration procedure—where there were several rounds of iterative planning and preregistration, and implementation and analysis—required modellers to be involved much earlier in the research process than usual. Changes in operational procedure and culture would need to occur at the case study institution to facilitate Adaptive Preregistration becoming commonplace	
⊖	**Preregistering flexible analyses is difficult and may delay the research process, especially as the analysis increases in complexity** Formulating alternative decision-pathways and their decision-triggers is challenging under certain conditions: when the underlying dataset is complex, unfamiliar to the modellers, zero-inflated, over-dispersed and small (low sample size). On the other hand, preregistration of straightforward projects would likely take less time due to increased familiarity with 'standard' or 'go-to' analyses for simpler data types and structures	The process of working through and documenting analysis decisions prior to analysing the data delayed completion of the analysis considerably when compared to expectation. The task of mentally forecasting potential analysis scenarios and decision pathways was particularly challenging in this instance, because the dataset was complex and unfamiliar to the modellers. Compared to knowledge and models of processes affecting fish responses to environmental flows, the vegetation models are relatively less developed. The lead modeller has much more familiarity modelling count data on fish and reflected that there would be less analytic unknowns in those instances, which would have made the task of preregistering flexibility much easier	*'Adaptive Preregistration did delay the research process initially, because we had to carefully consider all the likely scenarios to the end of the process and document those, as opposed to only needing to figure out the first part and then starting. I expected this delay to be about a month; however, the delays were much longer due to the fact that we had never tried preregistration before, we were also feeding back information into the Adaptive Preregistration process, and the model design and structure was very complex and different to things we had done in the past. It was a pretty challenging analysis—far more complex than many ecological studies I have worked on'*—CJ

(Continues)

TABLE 2 (Continued)

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕⊖	**Adaptive Preregistration is difficult at first and therefore slows the research process** but will become easier and more efficient with familiarity and experience with preregistration	The preregistration of decision trees and heuristics for triggering the selection of some models over others was the most difficult and time-consuming aspect of Adaptive Preregistration for the case study modellers. However, with practice and experience they expect Adaptive Preregistration will become easier and quicker	*'The process would definitely be easier and faster the next time, even for a project of the same complexity. Partly this is because of just knowing how it works and what to expect. The first one will always be the most difficult and slow. Going through the Adaptive Preregistration process once or twice will make it much more efficient'*—CJ
⊕	**Planning the analysis at the outset likely saves time in long run** Using a structured workflow that carefully considers modelling decision points and options in advance reduces scope for time-consuming back-and-forth adjustments later in the process. Having a documented plan to refer back to also saves time on similar, future tasks and report writing	Typically, the modelling team follows a more trial-and-error-based approach to model development, particularly for research and decision-making contexts where conceptual models, variable definitions and model parameters are uncertain (such as for the vegetation components of VEFMAP, compared to fish models). Adaptive Preregistration forced the modellers away from their usual iterative process towards carefully considered and well-justified a priori construction of the models. The case study modellers thought that the modelling process saved time in the coding of the models and by preventing iterative adjustments in the models and code and saving potential back-and-forth communications between modelling team members when deciding on and implementing and those changes	*'Undoubtedly using Adaptive Preregistration saved some time at the end of the process, but it is hard to say how much. Shifting the process from a more trial-and-error based approach to a very carefully considered construction of the model a priori, must surely have saved time in the coding and back-and-forth adjustments stages. The document was very helpful when it came to write the report. We also have a worked example that we are happy with that we can use to leverage off or remind ourselves how we approached a step in the past'*—CJ
⊕	**Preregistering model structures and predictors in advance strengthens links between methods and the original study aims/concepts**	Preregistration required careful consideration of the candidate predictors and possible model structures. The case study modellers acknowledged that while the broad approach to model development would likely have followed a similar progression (i.e. expanding on a basic Poisson model to address zero-inflation and overdispersion, while also reducing a candidate predictor set down to a maximal viable set), in the absence of Adaptive Preregistration, they may have proceeded without clear criteria for selecting one model over another (noting that these criteria were still relatively vague in the preregistered study)	*'If these models were developed without preregistration, much of this thinking would have been "on the fly," which may have resulted in better model fit (under a range of criteria) but perhaps weaker links to the study's aims and conceptual basis'*—JDY
⊕	**Adaptive Preregistration reduces researcher degrees of freedom** by shifting the nature of the modelling process away from a loosely bounded and iterative, trial-and-error style strategy towards a more constrained, goal-directed analysis. Incorporating loosely defined preliminary analyses with data subsetting facilitates data exploration being formally incorporated into the analytic decision-making process	When considering how the case study modelling would have occurred without being preregistered, both the project lead and lead modeller thought that the preregistration process led to similar outcomes to what would be expected under standard practices, but that the modelling process under Adaptive Preregistration may have resulted in a different set of predictors (including entirely different predictor variables) being selected because the model selection process was more constrained compared to usual practice	*'I do think we probably would have tried more variations on the model structure and variables without the preregistration. After starting with standard data checks to determine appropriate model structure, I would usually run simple models to see if I could get them to work. I would have then added complexity incrementally adjusting my model as I went and troubleshooting problems as they arose. I would have adjusted the definition of the regime variable, in particular, based on ecological expectations and model outputs (if required)'*—CJ
⊕	**Adaptive Preregistration increases research transparency** Specifying modelling and analysis plans in a preregistration together with justification of choices, ensures decisions are documented and reported in greater detail. Typically, only the final models are reported	The use of interim preregistrations of the Preliminary Analysis ensured that analyses of exploratory nature were documented when they otherwise would not have been, the results of which were expressly linked to candidate model structure selection in the preregistered Main Analysis. In particular, the project lead noted that early stages of the analysis were documented when typically, decisions and analyses at these early stages, such as data cleaning and processing decisions, would not have been reported	*'We would have conducted an extensive data review in the same way as with the preregistration, but the process would not have been as well documented and that would have made it much more likely that we would have had gaps in the process and therefore more potential for errors'*—CJ

TABLE 2 (Continued)

Type	Name and description	Case study example	Practitioner recommendations and reflections
⊕	**Adaptive Preregistration increases research accountability** Documenting the process helps demonstrate accountability to funders and decision makers	The project timeline included several reporting milestones and review processes. At early to middling timepoints, the modellers were able to share the preregistration document to funders to indicate how they intended to conduct the analysis, communicate rationale for particular analysis choices and demonstrate time investment	*'I was very glad to have the preregistration on hand to share or refer to for reporting milestones and review processes. The project was also long and complex with several delays, so it was nice to have a clear document outlining the process so that we could go back and check if we forgot an element of our plans'*—CJ
⊕	**Adaptive Preregistration encourages regular communication between modellers and others involved in the research** Clear communication among team members improves the chances that model outputs align with the project aims	*'The Adaptive Preregistration process (like static preregistration) requires clear communication between all people involved in collecting and modelling the data. Frequent communication provides more clarity about the purpose and scope of analyses, which may lead to model outputs being more closely aligned with project aims than would otherwise occur'*—JDY	
⊕	**Preregistration helps clarify good practice and process for less experienced researchers**	The lead and assistant modellers worked closely together during the case study, with the assistant modeller implementing part of the model development and coding of the models under the supervision of the lead modeller. The iterative nature of Adaptive Preregistration and its interim preregistrations could facilitate less experienced researchers to document their analysis choice rationale and validate this with experts or more experienced modellers before proceeding, so they could progress with confidence	*'The process helps to avoid future problems, improves transparency and quality of research practices, and is a very useful tool for less experienced researchers wanting to develop their skills. I strongly encourage everyone to go through it to understand how it works and how it could be of benefit to themselves and/or their colleagues'*—CJ
🚧	**Develop the preregistration and analysis over a short time period** so that institutional or organisational shifts do not disrupt the research process	*'The case study presented here was developed over several years, during which time there was staff turnover and changes in project constraints (including budgets). This complicated the process'*—JDY	
⊕	**Adaptive Preregistration provides a paper trail and archive of the current state of the project**, facilitating swift continuation of progress in case of unexpected loss of key personnel	*'It would have been extremely important to have the preregistration document on hand if something happened to our biometrician (e.g. sickness or other absence, which is common). If we had their code and the preregistration document, another biometrician would have been able to very quickly get everything they needed to progress things'*—CJ	

modelling and analysis activities, maintaining clear links between methodological decisions and the study aims and conceptual models while providing necessary flexibility to avoid committing to unsuitable modelling approaches (Table 2).

By facilitating data exploration in modelling through registered flexibility, data-subsetting and interim preregistrations, Adaptive Preregistration trades unbounded explorations for more goal-directed analyses, constraining researcher degrees of freedom and 'fishing expeditions'. In the absence of Adaptive Preregistration, the case study modellers felt they would have started with a much larger decision-space, trialling more variations on model structure and predictor variable operationalisations along the way, risking overfitting (Lewis et al., 2023).

While Adaptive Preregistration may increase overfitting risk compared to static preregistration due to its greater flexibility, it represents a substantial improvement over current practice (no preregistration) and includes specific safeguards against the QRPs that lead to overfitting. Through registered flexibility, Adaptive Preregistration prevents opportunistic QRPs, like 'model fishing', by making data-dependent decisions systematic rather than opportunistic, following transparent and prespecified criteria for modelling choices rather than optimising for desired outcomes (Gould, Fraser, Wintle, et al., 2026). Importantly, by formally incorporating exploratory activities into the preregistration process, Adaptive Preregistration exposes preliminary analyses that may involve QRPs but remain hidden in static preregistration approaches, where such exploration occurs before—and outside of—the formal preregistration. Our findings also show that Adaptive Preregistration increases both analytic and model transparency (Bodner et al., 2020; Lupia & Elman, 2014) by ensuring that modelling choices, steps, assumptions and expectations about outputs are articulated in greater detail than would typically occur (Schmolke et al., 2010). Documenting intermediate models, analyses and results exposes the influence of exploratory and data-contingent analyses on modelling choices. In our case study, the use of explicit prompts for choosing between and transparently justifying alternative model functional forms illuminated the analytic garden of forking paths, which too often remains opaque in ecological modelling (Fitzpatrick et al., 2021; Schmolke et al., 2010).

Modelling is considered an 'art form' as well as a science (Smaldino, 2020). Unpacking ingrained tacit knowledge may take time when writing preregistrations, especially the first time around, but there is clear value to capturing this expertise for future

researchers working on the same or similar problems. It took several attempts for our case study modellers to fully articulate registered flexibility when selecting model structure and functional form, perhaps because these procedures were second-nature in the modellers' practice, but by directing users to specific modelling tasks, preregistration prompted modellers to explicitly recognise, document and justify tacit modelling choices.

4.1 | When is Adaptive Preregistration useful?

We emphasise that Adaptive Preregistration is not intended to replace static preregistration for research that can be adequately served by existing static approaches. Rather, our case study represents a particularly challenging scenario for static preregistration: the modelling team was unfamiliar with the vegetation response data structure, had limited prior information about appropriate model forms for this system and faced institutional constraints requiring hypothesis-driven conclusions rather than purely open-ended exploratory analysis. These conditions—common in applied ecological research—made it impossible to specify a complete analysis plan a priori without risking commitment to unsuitable approaches. The case study exemplifies both information and procedural dependencies that Adaptive Preregistration is designed to address (Section 3.2): Preliminary Analysis results informed model structure choices (information dependency), and certain model types (zero-inflated Poisson, negative binomial) only became relevant after Preliminary Analysis revealed specific data properties (procedural dependency).

Whereas existing approaches to preregistration work well for familiar analytical contexts, we see Adaptive Preregistration as most beneficial when research involves these types of legitimate sequential decision-making and/or high data-model uncertainty, where researchers cannot reasonably predict whether intended models will be supported by unfamiliar or complex datasets. In such cases, static preregistration would force researchers into speculative commitments or require extensive post hoc explanations of deviations.

4.2 | Difficulties encountered using Adaptive Preregistration in our case study

Our case study researchers reported that the most challenging aspects of Adaptive Preregistration involved anticipating analysis decisions in advance, and changes to model management and timelines. Registering flexibility requires careful thought and time to adequately capture analytic decision-points and describe the process for resolving those decisions. In our case study, while the decision heuristic for choosing model functional form was accurately preregistered, the process for determining model structure was underspecified in the Main Analysis preregistration. This occurred because the preregistration did not specify that the heuristic should only apply to some models and not others (i.e. flow event models and not flow regime models, Figure 3).

The case study researchers reflected that anticipating potential analysis scenarios and decision pathways was complicated by large uncertainties about the underlying system dynamics in a complex and patchy dataset. The preregistration, while detailed, did not fully capture the modellers' intended and implemented model selection strategy, where after fitting the set of simpler models identified from the Preliminary Analysis (preregistration V2, Table 1), the modellers iteratively trialled different combinations of model terms to generate models closer to the more complex 'maximal' models identified in the initial preregistration (Table 1). The preregistered specification did not explain which combinations of model terms would be trialled, and in what order, rendering aspects of the model selection process unregistered. Project milestones with tight project turnaround times left limited time to iteratively review analysis specification decisions, resulting in some components of the modelling process being omitted from the preregistration. In an ideal world, following our Adaptive Preregistration user guide more closely may have avoided underspecificity in the case study preregistration, but our case study faced the same resource constraints typical of the real world, for example, costs required to upskill some members of the research team with git and GitHub.

Our second aim for the template and Adaptive Preregistration methodology was to facilitate 'preregistration checking'—the process of verifying that the analyses reported in a study matched what was preregistered (Pu et al., 2019). We approached preregistration checking through the Adaptive Preregistration methodology itself, specifically, through the registered flexibility and interim preregistrations mechanisms, together with our version-control protocol. These facilitated comparison of the data-dependent modelling pathways realised in the course of the study against the full set of decision alternatives at each decision-point, providing explicit rationale for any data-dependent preregistered analysis. By version-controlling the preregistration document, the genesis of the preregistration from one version of the next is made explicit, for example, GitHub's 'diff view' illustrates exactly how and where a document has changed between versions (Gould, Jones, et al., 2026; Appendix S3; Figure 1). Also, the use of GitHub issues in our protocol tracks the findings of each analysis and links them to the corresponding preregistration item. By viewing each preregistration item's GitHub issue, the implemented analyses and findings, together with any subsequent analyses, is clear to anyone checking the preregistration. Due to limitations in our own case study application (discussed above), the analysis paths and rationale were only transparent for some preregistration items (e.g. preregistered data cleaning and processing tasks, Figure 2.11, Appendix S7) or only partially transparent, with discussion of some changes to the analysis plan being recorded transparently (e.g. Figure 2.12, Appendix S7), but not explicitly linked to changes to the modelling and analysis code itself because changes to the modelling code were not tagged under the relevant GitHub issue. Consequently, we only partially achieved our aim of facilitating preregistration checking.

These aspects combined may result in researchers feeling like Adaptive Preregistration is slowing the research process because preregistration front-loads decision-making and planning in the project timeline, delaying analysis and implementation of the preregistration (Evans et al., 2023), which may be incongruent with existing institutional approaches to the governance, administration and operational support of modelling (i.e. 'model management' Arnold et al., 2020, p. 2). In hindsight, the case study preregistration process may have been more efficient and better specified if modellers and/or biostatisticians were brought earlier into the project's planning process to allow sufficient time to review, record and better connect analysis decisions to project aims. The case study was a particularly complex analysis with which to apply and test Adaptive Preregistration because of the data constraints and associated model structural uncertainty. For simpler analyses, Adaptive Preregistration would likely be less difficult once the process is broadly understood. We recognise that learning and implementing a new approach is unappealing at first, and although Adaptive Preregistration may save time later, it may require a rearrangement of project timelines and resources. Based on our reflections on the case study, we outline some advice for researchers wishing to implement Adaptive Preregistration in Box 1.

BOX 1 Tips for doing Adaptive Preregistration

Adaptive Preregistration should become easier and more efficient as researchers gain experience and familiarity with preregistration. When implementing Adaptive Preregistration for the first time, we recommend the following:

1. Start with simple descriptions or dot points in preregistration responses, building detail and complexity in analysis specification incrementally, as well as thorough and ongoing collaborative review.
2. Refer to existing preregistrations to guide researchers in completing their own. Although there are few examples in ecology now, with future uptake and an increasing number of journals offering the Registered Reports publication pathway (e.g. Conservation Biology, Biological Conservation, Journal of Ecology and Nature Ecology and Evolution), readers should expect an expanding pool of preregistrations to draw on. Existing registration repositories, such as OSF Registries (osf.io/registries/discover), can also be searched.
3. Developing refined templates for specific applications (e.g. prediction vs. inference, or specialised complex methods and study designs) may be necessary for adequately capturing analytical decisions. In contrast, with time and community uptake, a set of decision-points common across modelling methods and applications may naturally evolve over time through practice. Our template offers a starting point from which researchers can use, modify and add to as they see fit.
4. For some researchers, using version-control for transparently documenting the Adaptive Preregistration process may be too challenging. In such cases, or for simpler analyses, sequential analysis plans could be preregistered separately for the same study with metadata linking the chain of preregistrations without tracking associated analysis files. The OSF currently allows for preregistration updating after submission (but not updating of project files, help.osf.io/article/410-registration-files#ViewingPrevious).
5. Any first attempt to implement Adaptive Preregistration is unlikely to work perfectly, there will be mistakes and details that are omitted. Being upfront about this in study reporting is still better than avoiding preregistration entirely. See figure 1 in Lakens (2024) and table 1 in Willroth and Atherton (2024) for guidance on reporting preregistration deviations.

4.3 | Future work

While we believe we have partly addressed the problem of preregistering model-based research with the Adaptive Preregistration protocol, further work remains, particularly around determining the appropriate resolution of the preregistration template and developing cyberinfrastructure to effectively and transparently handle registering and reporting flexible analyses.

4.3.1 | What is the right resolution for a modelling preregistration template?

We aimed to design a preregistration template that was 'parsimonious', in that (1) template items should be specific and exhaustive enough to adequately constrain researcher degrees of freedom (Wicherts et al., 2016) so as to (2) facilitate transparent documentation of the modelling paths and decisions conducted during research, while (3) not requiring a prohibitive level of detail. Increasing the completeness and resolution of the template in capturing decisions throughout the modelling process may improve the ability of preregistration to restrict researcher degrees of freedom and transparently report results; however, it also makes the preregistration more challenging to complete. Refining the template during user testing greatly reduced the perceived difficulty of preregistration. However, our case study researchers conceded that some modellers may still consider the final template to be too cumbersome to invest the time in preregistration.

The writing and reporting phase of the case study also highlighted the need for ongoing refinement of the template content and structure in pursuit of our second use-case for the Adaptive Preregistration template: 'preregistration checking'. Registered flexibility for determining alternative model functional forms was spread over multiple subsections within the case study's preregistration template, hampering our ability to conduct preregistration checking

post-analysis (Table 2). We simplified the template during testing to ameliorate this issue, but it might be further improved by building modularity into the template and preregistration platform when preregistering multiple or complex models, such as in our case study.

We encourage researchers to preregister their research using our template and Adaptive Preregistration protocol flexibly, adopting elements that work and experimenting with alternative implementations that work better. New preregistration templates for expanding application of preregistration-specific modelling approaches and methods could be developed by integrating our template with existing reporting checklists and guidelines, such as ODMAP for reporting Species Distribution Models (Fitzpatrick et al., 2021). Alternately, researchers may identify a reduced version of our template with a minimum set of items that balances trade-offs between an exhaustive and specific template and the resource burden of prespecifying the analysis.

4.3.2 | Encouraging implementation and uptake of Adaptive Preregistration

Fostering wider adoption of (adaptive) preregistration will require coordinated action across the research ecosystem (O'Dea et al., 2021): researchers adopting and promoting preregistration as best practice, funders recognising and seeking out these quality standards, and journals encouraging uptake through publication preferences and recognition systems. Journals can support uptake by expanding Registered Reports programmes to accommodate Adaptive Preregistration, developing editorial guidelines for evaluating preregistered model-based studies and training reviewers to assess registered flexible analyses. Professional societies, like the Society for Open, Reliable, and Transparent Evolutionary Biology (SORTEE, https://sortee.org), have an important role in promoting best practices through training workshops and recognition schemes that celebrate transparent research practices. Engaging funders requires demonstrating that preregistration improves study quality, efficiency and reputation. Researchers can proactively market their preregistration practices as a quality standard when competing for funding, positioning themselves as the 'gold standard' for rigorous research.

The development of appropriate registries for Adaptive Preregistration is a critical barrier to address. Existing registries, such as the Open Science Framework (OSF, help.osf.io/article/330-welcome-to-registrations), OSF Registries (cos.io/products/osf-registries), As Predicted (aspredicted.org) and EcoEvoArXiv (ecoevorxiv.org), cannot accommodate Adaptive Preregistration, which requires dedicated instruments for capturing flexible analyses, reporting their outcomes and linking them to downstream decisions in subsequent preregistrations. To circumvent the inability to integrate Adaptive Preregistration into existing registries, we described and demonstrated in our case study a protocol for implementing Adaptive Preregistration and facilitating preregistration checking using git and GitHub. However, GitHub does not meet the minimum requirements for FAIR (Findable, Accessible, Interoperable and Reusable) data archiving (Pick et al., 2025), nor for preregistration (Haroz, 2022). Existing registries must be modified or new registries created to enable Adaptive Preregistration. Future registries could further streamline Adaptive Preregistration by displaying modular template content, shown to the user conditionally based on their preregistration responses, such that the preregistration 'expands' incrementally as items are completed. Publishers are well positioned to develop or commission (adaptive) preregistration registries that integrate with their submission systems, creating transparent linkages between the preregistration and final paper.

As open science practices are becoming increasingly embedded in the publication process, for example, reproducibility checklists at Conservation Biology and data- and code-checking in the peer-review workflow at Ecology Letters (Thrall et al., 2023), preregistration should further streamline the publication process by enhancing the transparency and rigour of the study. Beyond preregistration infrastructure, publishers could support this transition by developing standardised guidelines for reporting registrations in resulting manuscripts and providing clear editorial policies on how preregistered studies (that are not Registered Reports) should be handled during peer review. Such guidelines could address how to clearly reference preregistration items in manuscripts, report deviations transparently and present results from registered flexible analyses. For example, manuscripts should ideally reference the corresponding preregistration item number in the methods and results sections, explicitly acknowledging when and why any deviations occurred (e.g. see Gould et al., 2025 where deviations were reported with textboxes).

5 | CONCLUSIONS

Although challenges remain in preregistering ecological modelling, it is important to remember the costs and risks of not doing so. Ecological modelling has long been plagued by poor transparency. Incomplete reporting of the modelling process risks models being used in inappropriate applications or decisions and may mask poor model design or serious flaws in the model, resulting in adverse or irreversible outcomes when informing decision-making. Moreover, a lack of transparency in describing ecological models and reporting the modelling process provides significant opportunity for undisclosed researcher degrees of freedom, and hence, the possibility of QRPs, with researchers routinely executing alternative analyses and selectively reporting them when they do. Reproducible, transparent and reliable models are essential for sound conservation decision-making, and Adaptive Preregistration can improve documentation of modelling decisions, which is helpful for remembering and explaining the study's method at the time of report-writing, to aid others in understanding and replicating model-based research and to communicate to funders and stakeholders.

We provide a template and methodology for Adaptive Preregistration that we hope will extend the benefits of preregistration to model-based research in a way that strengthens rather than inhibits that research. Although we focus on ecological applications in this paper, many elements of the template are relevant across fields, and the open source template can be readily adapted to new fields and purposes.

AUTHOR CONTRIBUTIONS

Conceptualisation: Elliot Gould, Christopher S. Jones, Jian D. L. Yen, Hannah S. Fraser and Libby Rumpff. Data curation: Elliot Gould, Christopher S. Jones, Jian D. L. Yen and Henry F. Wootton. Formal analysis: Elliot Gould, Jian D. L. Yen and Henry Wootton. Funding acquisition: Elliot Gould, Libby Rumpff and Christopher S. Jones. Investigation: Elliot Gould, Christopher S. Jones, Jian D. L. Yen, Hannah S. Fraser, Megan K. Good, Cindy E. Hauser, Henry Wootton, David H. Duncan and Libby Rumpff. Methodology: Elliot Gould, Christopher S. Jones, Jian D. L. Yen, Hannah S. Fraser, Megan K. Good, David H. Duncan and Libby Rumpff. Software: Elliot Gould, Christopher S. Jones, Jian D. L. Yen and Henry Wootton. Supervision: Bonnie C. Wintle, Hannah S. Fraser and Libby Rumpff. Validation: Elliot Gould, Christopher S. Jones and Jian D. L. Yen. Visualisation: Elliot Gould, Christopher S. Jones, Jian D. L. Yen and Bonnie C. Wintle. Writing—original draft: Elliot Gould. Writing—review and editing: All authors.

ACKNOWLEDGEMENTS

Thank you to Lyndsey Vivian for assisting in data validation and cleaning in the VEFMAP case study. We also thank attendees for participation in our collaborative workshop. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australasian University Librarians.

FUNDING INFORMATION

This project was supported by the Arthur Rylah Institute for Environmental Research, Department of Energy, Environment and Climate Action of Victoria. Elliot Gould was supported by an Australian Government Research Training Program (RTP) Scholarship http://doi.org/10.82133/C42F-K220.

CONFLICT OF INTEREST STATEMENT

The authors declare no known conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70311.

DATA AVAILABILITY STATEMENT

Data are available via https://doi.org/10.5281/zenodo.19064000 (Jones, Yen, et al., 2026), https://doi.org/10.5281/zenodo.19064144 (Gould, Fraser, Jones, et al., 2026) and https://doi.org/10.17605/OSF.IO/SQ5J4 (Gould, Fraser, Wintle, et al., 2026).

ORCID

Elliot Gould https://orcid.org/0000-0002-6585-538X
Christopher S. Jones https://orcid.org/0000-0003-2833-0194
Jian D. L. Yen https://orcid.org/0000-0001-7964-923X
Hannah S. Fraser https://orcid.org/0000-0003-2443-4463
Henry F. Wootton https://orcid.org/0000-0001-6506-0248
Megan K. Good https://orcid.org/0000-0002-6908-1633
David H. Duncan https://orcid.org/0000-0003-4411-8214
Cindy E. Hauser https://orcid.org/0000-0001-5876-0342
Bonnie C. Wintle https://orcid.org/0000-0003-0236-6906
Libby Rumpff https://orcid.org/0000-0001-9400-8086

REFERENCES

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Alspaugh, S., Zokaei, N., Liu, A., Jin, C., & Hearst, M. A. (2019). Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

Arnold, T., Guillaume, J. H. A., Guillaume, J. H. A., Lahtinen, T. J., Lahtinen, T. J., Vervoort, W., & Vervoort, R. W. (2020). From ad-hoc modelling to strategic infrastructure: A manifesto for model management. *Environmental Modelling & Software*, *123*, 104563. https://doi.org/10.1016/j.envsoft.2019.104563

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128. https://doi.org/10.1016/j.ecolmodel.2013.11.009

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. T., & Wicherts, J. M. (2018). Ensuring the quality and specificity of preregistrations. *PsyArXiv [Preprint]*. https://doi.org/10.31234/osf.io/cdgyh

Baldwin, J. R., Pingault, J. B., Schoeler, T., Sallis, H. M., & Munafò, M. R. (2022). Protecting against researcher bias in secondary data analysis: Challenges and potential solutions. *European Journal of Epidemiology*, *37*(1), 1–10. https://doi.org/10.1007/s10654-021-00839-0

Bodner, K., Fortin, M.-J., & Molnár, P. K. (2020). Making predictive modelling ART: Accurate, reliable, and transparent. *Ecosphere*, *11*(6). https://doi.org/10.1002/ecs2.3160

Braga, P. H. P., Hébert, K., Hudgins, E. J., Scott, E. R., Edwards, B. P. M., Sánchez Reyes, L. L., Grainger, M. J., Foroughirad, V., Hillemann, F., Binley, A. D., Brookson, C. B., Gaynor, K. M., Shafiei Sabet, S., Güncan, A., Weierbach, H., Gomes, D. G. E., & Crystal-Ornelas, R. (2023). Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution. *Methods in Ecology and Evolution*, *14*(6), 1364–1380. https://doi.org/10.1111/2041-210X.14108

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). Do preregistration and Preanalysis plans reduce *p* -hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement. *Journal of Political Economy: Microeconomics*, *2*(3), 527–561. https://doi.org/10.1086/730455

Brudvig, L. A. (2017). Toward prediction in the restoration of biodiversity. *Journal of Applied Ecology*, *54*(4), 1013–1017. https://doi.org/10.1111/1365-2664.12940

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A.,

Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, *12*(4), 665–680. https://doi.org/10.1111/2041-210x.13559

Connolly, S. R., Keith, S. A., Colwell, R. K., & Rahbek, C. (2017). Process, mechanism, and modelling in macroecology. *Trends in Ecology & Evolution*, *32*(11), 835–844. https://doi.org/10.1016/j.tree.2017.08.011

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3). https://doi.org/10.1098/rsos.200805

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636–638. https://doi.org/10.1126/science.aaa9375

Evans, T. R., Branney, P., Clements, A., & Hatton, E. (2023). Improving evidence-based practice through preregistration of applied research: Barriers and recommendations. *Accountability in Research*, *30*(2), 88–108. https://doi.org/10.1080/08989621.2021.1969233

Fitzpatrick, M. C., Lachmuth, S., & Haydt, N. T. (2021). The ODMAP protocol: A new tool for standardized reporting that could revolutionize species distribution modelling. *Ecography*, *44*, 1067–1070. https://doi.org/10.1111/ecog.05700

Flournoy, J. C., Vijayakumar, N., Cheng, T. W., Cosme, D., Flannery, J. E., & Pfeifer, J. H. (2020). Improving practices and inferences in developmental cognitive neuroscience. *Developmental Cognitive Neuroscience*, *45*, 100807. https://doi.org/10.1016/j.dcn.2020.100807

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Gould, E. (2026). EcoConsModPreReg: A quarto template for preregistering modelling studies. *Zenodo*. https://doi.org/10.5281/zenodo.18907767

Gould, E., Fraser, H., Jones, C., Yen, J. D., Wootton, H., Duncan, D. H., Good, M., Hauser, C., Vivian, L., Wintle, B. C., & Rumpff, L. (2026). *Data, code and materials for "but I can't preregister my research": Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research*. https://doi.org/10.17605/OSF.IO/SQ5J4

Gould, E., Fraser, H. S., Parker, T. H., Nakagawa, S., Griffith, S. C., Vesk, P. A., Fidler, F., Hamilton, D. G., Abbey-Lee, R. N., Abbott, J. K., Aguirre, L. A., Alcaraz, C., Aloni, I., Altschul, D., Arekar, K., Atkins, J. W., Atkinson, J., Baker, C. M., Barrett, M., … Zitomer, R. A. (2025). Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, *23*(1), 35. https://doi.org/10.1186/s12915-024-02101-x

Gould, E., Fraser, H. S., Wintle, B. C., Rumpff, L., & Fidler, F. (2026). A framework for questionable research practices in ecological modelling. *EcoEvoArXiv [Preprint]*. https://doi.org/10.32942/X2DQ0K

Gould, E., Jones, C., Yen, J. D., Fraser Hannah, S., Wootton, H., Good, M., Duncan David, H., Rumpff, L., & Fidler, F. (2026). EcoConsPreReg: A guide to adaptive preregistration for model-based research in ecology and conservation. *Zenodo*. https://doi.org/10.5281/zenodo.19064144

Hämäläinen, R. P., & Lahtinen, T. J. (2016). Path dependence in operational research—How the modelling process can influence the results. *Operations Research Perspectives*, *3*, 14–20. https://doi.org/10.1016/j.orp.2016.03.001

Haroz, S. (2022). *Comparison of preregistration platforms*. OSF. https://doi.org/10.31222/osf.io/zry2u

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A. L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Hofman, J. M., Chatzimparmpas, A., Sharma, A., Watts, D. J., & Hullman, J. (2023). Pre-registration for predictive modelling (arXiv:2311.18807). *arXiv [Preprint]*. https://arxiv.org/abs/2311.18807

Ioannidis, J. P. A. (2022). Pre-registration of mathematical models. *Mathematical Biosciences*, *345*, 108782. https://doi.org/10.1016/j.mbs.2022.108782

Jones, C., Vivian, L., Wootton, H., Gould, E., Yen, J., Mole, B., Thomas, F., Backstrom, A., & Just, K. (2025). *Vegetation responses to managed river flow events and regimes* [Unpublished Report for the Department of Energy, Environment and Climate Action, Victoria]. Arthur Rylah Institute for Environmental Research.

Jones, C., Vivian, L., Wootton, H., Gould, E., Yen, J., Mole, B., Thomas, F., Sutton, F., Crowther, D., Backstrom, A., & Just, K. (2026). Vegetation responses to managed river flow events and regimes. *EcoEvoArXiv [Preprint]*. https://doi.org/10.32942/X2S67G

Jones, C. S., Yen, J. D. L., Gould, E., Wootton, H., & Vivian, L. (2026). Analysis code for: "Vegetation responses to managed river flow events and regimes" (0.21.2). *Zenodo*. https://doi.org/10.5281/zenodo.19064000

Kimmel, K., Avolio, M. L., & Ferraro, P. J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution*, *7*(9), 1525–1536. https://doi.org/10.1038/s41559-023-02144-3

Koivisto, E., & Mäntylä, E. (2024). Are Open Science instructions targeted to ecologists and evolutionary biologists sufficient? A literature review of guidelines and journal data policies. *Ecology and Evolution*, *14*(7), e11698. https://doi.org/10.1002/ece3.11698

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1), 117094. https://doi.org/10.1525/collabra.117094

Lewis, A. S. L., Rollinson, C. R., Allyn, A. J., Ashander, J., Brodie, S., Brookson, C. B., Collins, E., Dietze, M. C., Gallinat, A. S., Juvigny-Khenafou, N., Koren, G., McGlinn, D. J., Moustahfid, H., Peters, J. A., Record, N. R., Robbins, C. J., Tonkin, J., & Wardle, G. M. (2023). The power of forecasts to advance ecological theory. *Methods in Ecology and Evolution*, *14*(3), 746–756. https://doi.org/10.1111/2041-210x.13955

Liu, Y., Althoff, T., & Heer, J. (2020). Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. *Proceedings of the 2020 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3313831.3376533

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency. *PS: Political Science & Politics*, *47*(1), 19–42. https://doi.org/10.1017/s1049096513001716

MacEachern, S. N., & Van Zandt, T. (2019). Preregistration of modelling exercises may not Be useful. *Computational Brain & Behavior*, *2*(3–4), 179–182. https://doi.org/10.1007/s42113-019-00038-x

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. A. (2021). Both questionable and open research practices are prevalent in education research. *Educational Researcher*, *50*(8), 493–504. https://doi.org/10.3102/0013189X211001356

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G. J., Lavorel, S., Le Gall, L., Meslin, L., … Loreau, M. (2015). REVIEW: Predictive ecology

in a changing world. *Journal of Applied Ecology*, *52*(5), 1293–1310. https://doi.org/10.1111/1365-2664.12482

Neoh, M. J. Y., Carollo, A., Lee, A., & Esposito, G. (2023). Fifty years of research on questionable research practises in science: Quantitative analysis of co-citation patterns. *Royal Society Open Science*, *10*(10), 230677. https://doi.org/10.1098/rsos.230677

O'Dea, R. E., Parker, T. H., Chee, Y. E., Culina, A., Drobniak, S. M., Duncan, D. H., Fidler, F., Gould, E., Ihle, M., Kelly, C. D., Lagisz, M., Roche, D. G., Sánchez-Tójar, A., Wilkinson, D. P., Wintle, B. C., & Nakagawa, S. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*(1). https://doi.org/10.1186/s12915-021-01006-3

Parker, T., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology*, *33*(4), 747–750. https://doi.org/10.1111/cobi.13342

Pavelin, K., Cham, J. A., de Matos, P., Brooksbank, C., Cameron, G., & Steinbeck, C. (2012). Bioinformatics meets user-centred design: A perspective. *PLoS Computational Biology*, *8*(7), e1002554. https://doi.org/10.1371/journal.pcbi.1002554

Pick, J., Allen, B., Bachelot, B., Bairos-Novak, K., Brand, J., Class, B., Dallas, T., D'Amelio, P., Fenollosa, E., Fernández-Juricic, E., Gomes, D., Grainger, M., Guillemaud, T., John, C., Krasnow, R., Lagisz, M., Lequime, S., Maynard, D., Nakagawa, S., … Ivimey-Cook, E. R. (2025). The SORTEE guidelines for data and code quality control in ecology and evolutionary biology. *EcoEvoRxiv [Preprint]*. https://doi.org/10.32942/X24P8S

Popovic, G., Mason, T. J., Drobniak, S. M., Marques, T. A., Potts, J., Joo, R., Altwegg, R., Burns, C. C. I., McCarthy, M. A., Johnston, A., Nakagawa, S., McMillan, L., Devarajan, K., Taggart, P. L., Wunderlich, A., Mair, M. M., Martínez-Lanfranco, J. A., Lagisz, M., & Pottier, P. (2024). Four principles for improved statistical ecology. *Methods in Ecology and Evolution*, *15*, 2041-210X.14270. https://doi.org/10.1111/2041-210X.14270

Prosperi, M., Bian, J., Buchan, I. E., Koopman, J. S., Sperrin, M., & Wang, M. (2019). Raiders of the lost HARK: A reproducible inference framework for big data science. *Palgrave Communications*, *5*(1). https://doi.org/10.1057/s41599-019-0340-8

Pu, X., Zhu, L., Kay, M., & Conrad, F. (2019). Designing for preregistration in practice: Multiple norms and purposes. *Open Science Framework [Preprint]*. https://doi.org/10.31219/osf.io/w5ked

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 1. https://doi.org/10.5334/labphon.147

Schmolke, A., Thorbek, P., Deangelis, D. L., & Grimm, V. (2010). Ecological models supporting environmental decision making: A strategy for the future. *Trends in Ecology & Evolution*, *25*(8), 479–486. https://doi.org/10.1016/j.tree.2010.05.001

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., … Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology*, *51*(4), 207–218. https://doi.org/10.1027/1864-9335/a000425

Srivastava, S. (2018). Sound inference in complicated research: A multi-strategy approach. *PsyArXiv [Preprint]*. https://doi.org/10.31234/osf.io/bwr48

Thrall, P. H., Chase, J., Drake, J., Espuno, N., Hello, S., Ezenwa, V., Han, B., Mori, A., & Muller-Landau, H. (2023). From raw data to publication: Introducing data editing at Ecology Letters. *Ecology Letters*, *26*(6), 829–830. https://doi.org/10.1111/ele.14210

Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*(6), e03336. https://doi.org/10.1002/ecy.3336

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01832

Willroth, E. C., & Atherton, O. E. (2024). *Best laid plans: A guide to reporting preregistration deviations*. Association for Psychological Science. https://doi.org/10.1177/25152459231213802

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Zvereva, E. L., & Kozlov, M. V. (2021). Biases in ecological research: Attitudes of scientists and ways of control. *Scientific Reports*, *11*(1), 226. https://doi.org/10.1038/s41598-020-80677-4

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Description of the method for identifying the modelling workflow used in the workshop, accessible on the Open Science Framework at https://osf.io/tz5da/ (Gould, Jones, et al., 2026).

Appendix S2. Additional details of the user research and analysis process, including workshop materials and analysis, also available on the Open Science Framework at https://osf.io/sq5j4/ (Gould, Jones, et al., 2026).

Appendix S3. Detailed user guide to implementing adaptive preregistration transparently using git and GitHub. Reproduction of Gould, Fraser, Jones, et al. (2026) in .pdf format, also available on the Open Science Framework at https://osf.io/sq5j4/files/x2kpj (Gould, Jones, et al., 2026).

Appendix S4. Preregistration template in .pdf format, generated from the quarto markdown extension 'EcoConsModPreReg'. Available on GitHub at https://github.com/egouldo/EcoConsModPreReg (Gould, 2026) and on the Open Science Framework at https://osf.io/sq5j4/files/gxkdz (Gould, Jones, et al., 2026).

Appendix S5. Preregistration template in .docx format, from the quarto markdown extension 'EcoConsModPreReg'. Available on GitHub at https://github.com/egouldo/EcoConsModPreReg (Gould, 2026) and on the Open Science Framework at https://osf.io/sq5j4/files/m6452 (Gould, Jones, et al., 2026).

Appendix S6. Final version of the case study preregistration. Available on GitHub at https://github.com/egouldo/VEFMAP_VEG_Stage6/blob/f10b03d48ff3107387262882ff980072f0fb3c06/analysis/preregistration/preregistration.md (Jones, Yen, et al., 2026) and on the Open Science Framework at https://osf.io/sq5j4/files/wnc4s (Gould, Jones, et al., 2026).

Appendix S7. Summary report of the case study Preliminary Analysis, accessible on the Open Science Framework at https://osf.io/sq5j4/files/zbr92 (Gould, Jones, et al., 2026).

How to cite this article: Gould, E., Jones, C. S., Yen, J. D. L., Fraser, H. S., Wootton, H. F., Good, M. K., Duncan, D. H., Hauser, C. E., Wintle, B. C., & Rumpff, L. (2026). 'But I can't preregister my research': Improving the reproducibility and transparency of ecology and conservation with adaptive preregistration for model-based research. *Methods in Ecology and Evolution*, *00*, 1–20. https://doi.org/10.1111/2041-210x.70311