Towards improved rainfall-runoff modelling in changing climatic conditions

Keirnan James Andrew Fowler

Open Researcher and Contributor ID (ORCID): 0000-0002-1983-0253

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

November 2017

Hydrology and Water Resources Group
Department of Infrastructure Engineering
The University of Melbourne
Australia
Declaration

This is to certify that:

1. this thesis comprises only my original work towards the degree of Doctor of Philosophy,

2. due acknowledgement has been made in the text to all other material used,

3. the thesis is fewer than 100,000 words in length exclusive of tables, maps, bibliographies and appendices.

Keirnan Fowler
Melbourne, November 2017
Rainfall-runoff models are useful tools in water resource planning under climate change. They are commonly used to quantify the impact of changes in climatic variables, such as rainfall, on water availability for human consumption or environmental needs. Many parts of the world are likely to see changes in future climate, and some regions are projected to be substantially drier, possibly with threatened water resources. Given the importance of water to the economy, environment, geopolitical stability and social wellbeing, reliable tools for understanding future water availability are vital.

However, literature would suggest that the current generation of rainfall-runoff models are not reliable when applied in changing climatic conditions. Simulations of historic case studies such as the Millennium Drought in South East Australia indicate that models often perform poorly, underestimating the sensitivity of runoff to a given change in precipitation. Many hydrologists have assumed that these deficiencies are due to the model structures themselves - that is, the underlying model equations. However, it is possible that the explanation is broader, and can only be understood via holistic approaches that examine the entire modelling process. This research, presented in four parts, aims to understand and improve various elements of this process.

Part 1 (Chapter 4) investigates whether poor model performance is due to insufficient model calibration and evaluation techniques. An approach based on Pareto optimality is used to explore trade-offs between model performance in different climatic conditions. Five conceptual rainfall-runoff model structures are tested in 86 catchments in Australia. Comparison of Pareto results with a commonly used calibration method reveals that the latter often misses potentially promising parameter sets within a given model structure, giving a false negative impression of the capabilities of the model. This suggests that existing model structures may be more capable under changing climatic conditions than previously thought. The aim of Part 1 is to critically assess commonly used methods of model calibration and evaluation, rather
than to develop an alternative calibration strategy. The results indicate that caution is needed when interpreting the results of differential split sample tests.

Having demonstrated deficiencies in commonly used calibration methods, Parts 2 and 3 (Chapters 5 and 6) examine alternative calibration strategies. The aim is to identify calibration metrics capable of finding parameter sets with robust performance, even if climatic conditions change compared to the calibration period. Part 2 follows a three-part process to identify which metrics (if any) can identify the robust parameter sets using pre-change data only. The three parts are: randomly generating a large ensemble of parameter sets; identifying parameter sets in the ensemble that provide robust simulations both before and after a change (drying) in climatic conditions; and calculating multiple performance metrics for each ensemble member. Traditional objective functions are trialled, along with less common indices such as the degree of replication of observed hydrologic signatures. The most promising metrics are then tested more rigourously in Part 3, which uses guided search algorithms selected in accordance with metric type (objective function or hydrologic signature), including: calibration by matching of hydrologic signatures (using the DREAM-ABC algorithm), optimisation of global objective functions (using the CMA-ES algorithm), and hybrid approaches blending global objective functions with signatures (using the Pareto approach AMALGAM). The results indicate considerable scope for improved calibration, relative to commonly used approaches. Metrics that consider dynamics over a variety of timescales (e.g. annual, not just daily) are more promising, as are objective functions using the sum of absolute errors rather than the sum of squared errors. The key recommendations of Part 2 and 3 are to avoid ‘least squares’ approaches (such as optimising the NSE, RMSE and similar approaches like the KGE) and adopt sum of absolute error and/or metrics considering a variety of timescales, wherever simulations of a drying climate are required.

Parts 1-3 confirm the importance of calibration methods when modelling under changing climates. This raises the question: in what circumstances should the focus be on improving calibration methods versus improving model structures, or alternatively on other issues such as poor data quality? Although recent literature
has presented various tools for model evaluation - usually using variants of the Differential Split Sample Test (DSST) - there is less focus on such questions. Thus, a modeller whose model has failed the DSST is largely without guidance as to next steps.

Part 4 (Chapter 7) provides guidance for this question within a framework based on Pareto optimality. Similarly to Part 1, modelling objectives are set over multiple historic periods with contrasting climatic conditions. The framework allows cases of DSST failure to be categorised as either: (a) cases of model structural failure, where no parameter set in a model structure can meet all modelling objectives in all periods, indicating the need for structural changes or improved data; or (b) cases where modelling objectives are attainable by the model structure, but the DSST calibration method failed to find the right parameter set(s). The framework outlines separate steps to follow for each of the above categories. Many steps in the framework can be populated by existing sensitivity analysis techniques, but new techniques are designed for some steps, such as the diagnosis of structural inadequacies by analysis of ‘drift’ in hydrologic signature error as climatic conditions change. The framework is demonstrated using a case study from Australia and the IHACRES model structure. Limitations of inferring future hydrologic processes from historic data are also discussed.

This research underscores the joint importance of model structures and calibration methods when modelling changing climatic conditions, providing practical guidance for holistic improvement of the modelling process. By prompting more credible runoff projections, it is hoped that this research leads to more robust decisions that safeguard the future of water resources for people and our planet.
Acknowledgements

Firstly to my principal supervisor and friend Murray Peel, for many long conversations, good humour and for graciously supporting me to follow whatever research enquiries seemed most engaging. Murray tirelessly prioritises time for his students and advocates on their behalf. Also to Murray’s wife, Fiona, for good company and excellent advice on the finer points of art and history.

To Andrew Western for seeing through the fog and being able to express ideas with clarity, for maintaining an open door policy despite his many hats, and for always keeping the big picture in mind.

To Lu Zhang (CSIRO), for patiently and generously enduring so many phone hookups in which he was the only person off-site (and associated technological glitches), for prompting countless improvements to my written work through well-placed commentary and questions, and for extending hospitality to me while in Canberra.

Many others at the University of Melbourne contributed significantly to this work. The kind assistance and good company of Tim Peterson is gratefully acknowledged, particularly with regards to technical matters of hydrology and computing. Also thanks to Tim for providing the \LaTeX{} template for this document (with considerable help and additions by Michael Rigby). The kind and patient assistance of Lev Lafayette and Daniel Tosello was instrumental in facilitating analyses using the University’s High Performance Computer. The good-humoured administrative assistance of Pauline Woolcock is gratefully acknowledged. Conversations with and assistance from Rory Nathan, Avril Horne and Tom McMahon were also much appreciated.

Part of this research was completed while visiting the University of Bristol, UK. The support and hospitality of Jim Freer, Ross Woods and Thorsten Wagener are gratefully acknowledged. All three were very generous with their support, particularly Thorsten through supporting my stay administratively, Ross through incisive and thoughtful technical feedback, and Jim through generously long conversations.
(some of them in the great outdoors) and excellent gin. The stay in Bristol was made so much more pleasant by interaction with many in the Bristol research group, particularly the excellent hospitality and technical feedback of Wouter Knoben, and through the encouragement and tireless assistance of Gemma Coxon.

The financial support of the Australian Government is gratefully acknowledged, particularly through an Australian Postgraduate Award, and also via Murray Peel’s Future Fellowship (FT120100130). Researchers at the University of Bristol are supported by the UK’s Natural Environment Research Council grant MaR-IUS: Managing the Risks, Impacts and Uncertainties of droughts and water Scarcity (NE/L010399/1). Also, thanks to the Rotary Club of Balwyn, Victoria, whose financial assistance enabled part of this research to be presented at the 2016 AGU conference in San Francisco, USA.

Thanks to Robert Morden for mentoring and advice that honed my technical skills early in my career, for continued friendship, and for assistance with data procurement for this thesis. Thanks to my fellow PhD students and officemates for their support - Alex, Rubel, Mukta, Emma, Amir, Danuta, Anita, Julie and many others. To Margarita, Sina and Tash for interesting and productive conversations, and to Mike and Sam for so many lunch time chats and continual encouragement.

To my parents and sister: thanks for your continued support, love and blessing in my life, particularly over the four years of this PhD. To Kelly: thank you for walking on this journey with me and still, somehow, wanting to marry me at the end.

Finally, to the One who animates everything and gives us a mind to study it, and to whom no acknowledgement ever suffices.
Preface

The research described in this thesis is entirely my own work, and the contributions of others are limited to supervision, technical assistance and/or manuscript review. No part of this thesis has been submitted for another qualification, and no part was carried out prior to PhD enrolment. Please see the previous page for acknowledgement of funding sources.

Chapters 4 to 7 are each formatted as journal articles and have been submitted or accepted as follows:

- Chapter 4 has been published as Fowler, K., Peel, M., Western, A., Zhang, L., and Peterson, T. J., 2016: Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, Water Resources Research, 52, doi:10.1002/2015WR018068.

- Chapter 5 has been submitted as Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved model calibration under changing climatic conditions, 1: an exploratory analysis. In revision following peer review by Water Resources Research.

- Chapter 6 has been submitted as Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved model calibration under changing climatic conditions, 2: comparison of methods. In revision following peer review by Water Resources Research.


This pdf document contains imbedded links to aid navigation. The links cover sections, cited references, figures, and tables. To view a cited reference, click the year
of the reference, or if the author appears in italics, the author name is also included in the link. In the reference list, each reference is followed by a list of pages on which the reference is cited, allowing easy navigation back to the body of the thesis.
Contents

1 Introduction .......................................................... 1
  1.1 Problem statement .............................................. 1
  1.2 Aim, research questions and scope .......................... 2
  1.3 Thesis structure ............................................... 4

2 Literature Review .................................................. 5
  2.1 Introduction ................................................... 5
     2.1.1 Climatic variability and change ...................... 5
     2.1.2 Water resource planning for changing climatic conditions 7
     2.1.3 What is hydrological projection? ...................... 8
     2.1.4 Approaches to hydrological projection ................. 9
     2.1.5 Concepts and definitions in rainfall-runoff modelling 10
     2.1.6 Introduction to study area ............................ 11
  2.2 Problems with modelling in changing climate ............ 14
     2.2.1 The differential split sample test ................... 14
     2.2.2 Reported differential split sample test results .... 15
     2.2.3 Nonstationary parameters .............................. 16
     2.2.4 Characterising deficiencies ............................ 18
     2.2.5 Possible causes for deficiencies ..................... 19
     2.2.6 Summary ................................................. 19
  2.3 Calibration of daily rainfall-runoff models ............. 20
     2.3.1 Manual calibration ..................................... 20
     2.3.2 Automatic calibration .................................. 21
     2.3.3 Choosing an objective function ...................... 22
     2.3.4 Consideration of uncertainty ......................... 23
     2.3.5 Ensemble methods ..................................... 25
     2.3.6 Multi-objective calibration and hydrologic signatures 27
     2.3.7 Calibration methods contributing to existing literature consensus 30
     2.3.8 Summary of findings for review of calibration methods 32
  2.4 Improving rainfall-runoff model structures ............. 33
     2.4.1 Overview of physically based and conceptual modelling 33
2.4.2 Inferring structural deficiencies from simulation results ..... 36
2.4.3 Possible structural deficiencies .......................... 37

3 Research Questions

3.1 Potential research questions identified in Literature Review ....... 45
3.2 Prioritisation of research questions ............................. 46
3.3 Final list of research questions .................................. 48

4 Revisiting model deficiency in changing climate

4.1 Chapter Summary .................................................. 51
4.2 Introduction ......................................................... 52
4.3 Method .............................................................. 58
  4.3.1 Rationale ....................................................... 58
  4.3.2 Multi-criteria analysis and Pareto search method .......... 61
  4.3.3 Rainfall-runoff model structures ............................ 62
  4.3.4 Study area .................................................... 64
  4.3.5 Input data ..................................................... 66
  4.3.6 Defining dry periods and wet periods .................... 67
  4.3.7 Objective functions ......................................... 70
  4.3.8 Results checking ............................................ 71
4.4 Results ............................................................ 73
  4.4.1 Performance when optimising to each objective in isolation . 73
  4.4.2 Pareto Curve results ......................................... 74
  4.4.3 Identifying model structures that meet modelling standards . 79
  4.4.4 Examination of catchments where models failed .......... 83
4.5 Discussion .......................................................... 87
  4.5.1 Results summary .............................................. 87
  4.5.2 Getting the right answers for the wrong reasons? .......... 88
  4.5.3 The role of data errors ...................................... 89
  4.5.4 Relevance to future model improvements ................ 91
  4.5.5 Minimising false negatives ................................ 92
  4.5.6 Climate change: beyond the scope of historical observations? 93
  4.5.7 Research challenges ......................................... 93
4.6 Conclusions and recommendations ............................... 95

5 Improving calibration: exploratory analysis

5.1 Chapter Summary .................................................. 99
5.2 Introduction ......................................................... 100
5.3 Method .............................................................. 105
  5.3.1 Method overview ........................................... 106
  5.3.2 Metric selection and definition ............................ 113
CONTENTS

7.4 Case study ............................................. 196
  7.4.1 Example catchment and data ....................... 196
  7.4.2 Example model structures ....................... 198
  7.4.3 Application of framework ....................... 199
  7.4.4 Choose acceptance metrics and thresholds (Step 1a) ...... 199
  7.4.5 Define periods (Step 1b) ....................... 199
  7.4.6 Choose calibration method (Step 1c) .................. 200
  7.4.7 Model calibration/evaluation (Step 2a & 2b) ............ 200
  7.4.8 Pareto test (Step 3) .......................... 201
  7.4.9 Analyse change (Step 4a) ...................... 201
  7.4.10 Analyse model structure (Step 4b) ............... 203
  7.4.11 Model structural change (Step 4c) ............... 209
  7.4.12 Repeat model calibration/evaluation (Step 2a & 2b) .... 209
  7.4.13 Repeat Pareto test (Step 3) .................. 210
  7.4.14 Revise calibration methods and repeat DSST (Step 5a) .... 211
  7.4.15 Additional data (Step 5b) ....................... 212
  7.4.16 Summary of case study application .................. 213

7.5 Discussion ........................................... 214
  7.5.1 Interpretation of the DSST ....................... 214
  7.5.2 Further research in calibration methods ............... 214
  7.5.3 Are models that pass the DSST ‘adequate’? .............. 215
  7.5.4 Generality of the framework .................... 217

7.6 Conclusions and recommendations ....................... 218

8 Discussion and Conclusions .......................... 219
  8.1 Research overview ................................ 219
  8.2 Limitations and assumptions ....................... 221
  8.3 Future research directions ....................... 224
    8.3.1 Better characterisation of model deficiencies ........ 224
    8.3.2 Improved model structures, using the RQ3 framework .... 225
    8.3.3 Improved calibration methods .................. 226
    8.3.4 Other ........................................ 227
  8.4 Original research contributions ................... 228
  8.5 Concluding research statement ................... 229

Appendix A Study catchment details .................. 231

Appendix B Supplementary material for Chapter 4 .... 249
  B.1 Introduction ...................................... 249
  B.2 Categorisation of model failure based on shape of the Pareto Curve .... 251
  B.3 Discussion of bushfires .......................... 252
CONTENTS

B.4 Selected figures and tables .............................................. 253

Appendix C Supplementary material for Chapter 5 .............................. 265
   C.1 Introduction ...................................................... 265
   C.2 Combined ECDFs for each model structure ......................... 266
   C.3 Method description for two-metric 'Flow Duration Curve' signature 268
   C.4 Selected figures and tables ...................................... 269

Appendix D Supplementary material for Chapter 6 .............................. 277
   D.1 Introduction ...................................................... 277
   D.2 DREAM-ABC settings (with associated analysis) ................. 278
      D.2.1 Overview of DREAM-ABC settings ........................... 278
      D.2.2 Analysis supporting adopted settings ...................... 279
   D.3 Prior study of signature combinations ............................ 281
   D.4 DREAM-ABC case study selection rules ......................... 286
   D.5 Hybrid method expanded results ................................ 291
   D.6 Data depth performance review ................................ 292
   D.7 Results using alternative evaluation metrics ..................... 296

Appendix E Supplementary material for Chapter 7 .............................. 303
   E.1 Introduction ...................................................... 303
   E.2 Step 4a notes for case study catchment ........................... 304
   E.3 Additional Step 4b analysis .................................... 306
   E.4 General application of Step 5b ................................. 308
   E.5 Application of framework to multiple catchments ................ 310
   E.6 Selected tables and figures .................................... 311

References ................................................................. 315
## List of Figures

2.1 Map of study region showing catchment locations and properties .......................... 12
4.1 Explanatory diagram of DSST outcomes using idealised categories ..................... 59
4.2 Illustrative hypothetical Pareto Curves .......................................................... 60
4.3 Catchment properties for the 86 study catchments ............................................. 69
4.4 Characteristics of 7-year dry periods ............................................................. 70
4.5 KGE results when optimised to only one objective ........................................... 75
4.6 Scatter plots of single objective results ......................................................... 77
4.7 Example Pareto front for IHACRES applied to the Rocky River ....................... 78
4.8 Pareto results for all model structures applied to the Rocky River ................... 78
4.9 Example showing modelling standards definition ............................................. 81
4.10 Modelling standards met by single-objective DSST and AMALGAM ............... 83
4.11 Catchment characteristics where model structures failed standards ............... 86
4.12 Long-term moving average simulation bias for a GR4JMOD example ............... 89

5.1 Exploratory analysis method for Chapter 5, 1: signatures ............................... 108
5.2 Exploratory analysis method for Chapter 5, 1: obj. functions .......................... 112
5.3 Boxplots of adjusted KSD values for each hydrologic signature ...................... 129
5.4 Boxplots of adjusted KSD values for each objective function ......................... 131
5.5 Gridded performance of each objective function ........................................... 132

6.1 DSST results for objective functions .......................................................... 161
6.2 Example results for hybrid method .............................................................. 163
6.3 DSST results for hybrid method ................................................................. 164
6.4 Example results for ABC ............................................................................... 166
6.5 DSST results for ABC ................................................................................... 168
6.6 DSST results - comparison of methods ......................................................... 170
6.7 DSST results - scatter plot comparisons with reference method ..................... 171
6.8 Accumulation of squared and absolute errors with time ................................. 174
6.9 Example year-by-year performance for Split KGE ........................................ 175

7.1 Flow chart for model improvement framework ................................................ 190
LIST OF FIGURES

7.2 Map and historical hydrologic data for study catchment 197
7.3 Pareto testing of IHACRES-A, showing acceptance thresholds 202
7.4 Analysis of climatic ‘drift’ in hydrologic signature error 206
7.5 Pareto testing of IHACRES-B, with signature analysis 210
7.6 Comparison of split sample results for different calibration methods 213
7.7 Conceptual levels of model adequacy for changing climate 216

B.1 Details of model failure for each model structure 251
B.2 All Pareto curves derived for Chapter 4, 1 of 2 254
B.3 All Pareto curves derived for Chapter 4, 2 of 2 255
B.4 KGE and NSE comparisons 256
B.5 Graphical comparison of AMALGAM results with CMA-ES 257
B.6 Selected cases of divergent Pareto curves with similar endpoints 258
B.7 Map of catchments where no model structure could meet standards 259
B.8 Long-term simulation bias for selected parameter sets, 1 of 4 260
B.9 Long-term simulation bias for selected parameter sets, 2 of 4 261
B.10 Long-term simulation bias for selected parameter sets, 3 of 4 262
B.11 Long-term simulation bias for selected parameter sets, 4 of 4 263

C.1 Combined ECDFs for each model structure, for hydrologic signatures 267
C.2 Combined ECDFs for each model structure, for objective functions 268
C.3 Histograms of observed signature values across 221 catchments 274
C.4 Comparison of LHS random samples with Pareto Fronts 275

D.1 Gelman Rubin statistic versus DREAM-ABC behavioural population 282
D.2 Number of DREAM-ABC chains versus volume of parameter space 283
D.3 Effect of $\epsilon$ value on the volume of parameter space 284
D.4 Hybrid method expanded results for Chapter 6, 1 of 2 293
D.5 Hybrid method expanded results for Chapter 6, 1 of 2 294
D.6 Data depth performance review 295
D.7 Chapter 6 results using NSE as evaluation metric 297
D.8 Chapter 6 results using model bias as evaluation metric 298
D.9 Chapter 6 results using Linear Correlation as evaluation metric 299
D.10 Chapter 6 results using model variability as evaluation metric 300
D.11 Chapter 6 results using NSE$_{LF}$ as evaluation metric 301

E.1 Rainfall double mass curves for catchment 613002 305
E.2 Two additional analyses for Step 4a 308
E.3 Framework results for 86 catchments 311
E.4 Simulation dynamics for IHACRES-A and B 312
List of Tables

4.1 Details of the five conceptual rainfall-runoff model structures tested . 64
4.2 Statistical tests on catchment characteristics, grouped by Pareto results 86

5.1 Objective function descriptions for Chapter 5 . . . . . . . . . . . . . . . 117
5.2 Signature descriptions for Chapter 5 . . . . . . . . . . . . . . . . . . . . . 119

A.1 Selection of catchments from Hydrologic Reference Stations . . . . . 231

B.1 Characteristics of catchments where all models failed . . . . . . . . . . 253

C.1 Case studies accepted for analysis in Chapter 5 . . . . . . . . . . . . . . 270

D.1 Signature combination prior study results (1 of 2) . . . . . . . . . . . . . 286
D.2 Signature combination prior study results (2 of 2) . . . . . . . . . . . . . 287
D.3 DREAM-ABC selected case studies for Chapter 6 . . . . . . . . . . . . . 288

E.1 Parameter list for IHACRES-A and IHACRES-B . . . . . . . . . . . . . . 313
E.2 Possible points of investigation for Step 4a . . . . . . . . . . . . . . . . . 314
Chapter 1

Introduction

1.1 Problem statement

Climate change is likely to cause changes in water availability in many regions of the world (Dai, 2011; Trenberth, 2011). Although increased precipitation is expected in the tropics, climate projections indicate reduced precipitation in many places including southern Australia (Chiew et al., 2009), the Mediterranean region (Gao and Giorgi, 2008), subtropical North America (U.S. Global Change Research Program, 2009), and parts of Africa (Park Williams and Funk, 2011). These reductions may be accompanied by changes in rainfall variability (de Boer, 2009), rainfall seasonality (Weiss et al., 2009), and snow fraction (Barnett et al., 2005), in addition to possible trends in evaporation and transpiration associated with a warming world (Donohue et al., 2010). Although climate projections remain uncertain, in many regions there is strong consensus among models about direction of change (eg. Chiew et al., 2009; Faramarzi et al., 2013; Addor et al., 2014).

This has important implications for society and the environment. Water is a basic human need, and economic growth is a ‘surprisingly thirsty business’ (World Bank, 2016). Changing water availability may cause geopolitical instability and potential conflict (Raleigh and Kniveton, 2012; Hsiang et al., 2013). Ecosystems may incur species loss and environmental degradation (Hughes, 2003; Hobday and Lough, 2011). Adaptation may be costly, requiring new or retrofitted infrastructure and
difficult choices between economic, environmental and social uses of water (van Dijk et al., 2013; Aghakouchak et al., 2014). These issues, among others, have motivated the International Association of Hydrologic Sciences to call for scientific focus on ‘change in hydrology and society’ for the decade 2013-2022 (Montanari et al., 2013).

Reliable tools for understanding future water availability are vital. Tools relevant to future water planning include: (i) global climate models (GCMs), which estimate precipitation and temperature under future greenhouse emissions scenarios; (ii) rainfall-runoff models, which estimate runoff resulting from GCM outputs; and (iii) water systems models, which investigate various options for operation of infrastructure and associated trade-offs, given projected runoff.

The second category, rainfall-runoff models, is the focus of this research. Recent literature suggests the current generation of rainfall-runoff models are not reliable in changing climate. The models often perform poorly when simulating historic droughts, underestimating the sensitivity of runoff to a given change in precipitation (Vaze et al., 2011; Coron et al., 2012, 2014; Saft et al., 2016a). Models that cannot simulate historic droughts are unlikely to be adequate to simulate runoff for future projections, which are often beyond the envelope of historic climatic variability (Meehl et al., 2007). This represents a significant impediment to provision of credible runoff projections for water resources management.

1.2 Aim, research questions and scope

This research focusses on improving runoff simulations in changing climatic conditions. The broad aims are twofold: to investigate causes of poor performance of rainfall-runoff models over changing climatic conditions; and to formulate changes to modelling practice that may improve performance of rainfall-runoff models over changing climatic conditions.

More specifically, these aims are developed into the following research questions. The second and third questions are informed by results from preceding questions.
1.2. AIM, RESEARCH QUESTIONS AND SCOPE

1. Is the poor performance of rainfall-runoff models over historic droughts due to poor calibration methods, model structural deficiencies, or both?

2. Which calibration metrics result in simulations that are robust to reductions in long-term average precipitation?

3. In cases where a model has already failed split sample testing, what general steps should be followed to improve simulations under changing climates?

The scope of the research follows directly from the problem statement:

a **Model type:** The focus is on conceptual rainfall-runoff models because they are typically used in climate projection studies. Concepts from physically based modelling are within scope provided they are deemed relevant to diagnosing problems with conceptual models. Runoff projections are usually concerned with water quantity, so water quality models are not in focus.

b **Data used in modelling:** Models used for runoff projections are typically forced only by rainfall and potential evapotranspiration, and calibrated only to streamflow, if available. Consideration of other data types (eg. groundwater, remote sensing) are within scope if deemed relevant to diagnosing problems with conceptual models, or improving simulations under changing climates.

c **Choice of catchments:** The literature consensus (that rainfall-runoff models are not reliable in changing climate) is based on studies in unregulated catchments. Thus, regulated catchments are not in focus. Flow in unregulated catchments may still be significantly altered by human activities, and consideration of such impacts is within scope.

Lastly, the geographic scope is limited to a single region, namely southern and eastern Australia, which has been subject to recent long droughts that provide case studies of apparent model failure.
1.3 Thesis structure

This thesis is structured as follows. Chapter 2 reviews relevant literature and Chapter 3 lists perceived research gaps, describes how these are prioritised, and restates the research questions. Chapters 4 - 7 describe the research conducted. Note that each of these chapters has been formulated as a stand-alone research article, as described in the preface. Research Question 1 is addressed in Chapter 4, while Research Question 2 is spread between Chapters 5 and 6. Research Question 3 is addressed in Chapter 7. Chapter 8 discusses the findings and limitations of the PhD project as a whole and summarises the major contributions and conclusions of this research.
Chapter 2

Literature Review

This literature review is structured in four sections. Section 2.1 serves as an introduction, discussing the context, defining key terms and concepts such as projection and simulation, and describing the study area. Section 2.2 reviews literature indicating deficiencies in rainfall-runoff models when applied in changing climate. Section 2.3 discusses methods to calibrate daily rainfall-runoff models. Section 2.4 reviews possible model structural deficiencies in the context of changing climate. This chapter should be read together with Chapter 3, which summarises the perceived research gaps and explains the adopted research questions.

2.1 Introduction

2.1.1 Climatic variability and change

Natural systems are inherently variable on many timescales. These include familiar short term daily and seasonal cycles, through to multi-decadal and longer variability (Hurst, 1951; Kiem, 2004; Cohn and Lins, 2005; Koutsoyiannis, 2011; Cook et al., 2016; O’Connell et al., 2016). Cycles influencing climatic conditions on scales of years and decades include the El Nino Southern Oscillation (ENSO), Interdecadal Pacific Oscillation (IPO), Indian Ocean Dipole (IOD) and Southern Annular Mode (SAM) (Power et al., 1999; Saji et al., 1999; Mantua and Hare, 2002; Thompson and Solomon, 2002; Peel et al., 2002; Murphy and Timbal, 2008; Verdon-Kidd and
CHAPTER 2. LITERATURE REVIEW

Kiern, 2009; Henley et al., 2011; Gallant et al., 2012). Due to irregular periodicity and interactions among these cycles, dry or wet conditions may persist for many years (Power et al., 1999; Kiern, 2004). In some cases, unexpected shifts may also be caused by feedbacks within the hydrometeorological system (D’Odorico and Porporato, 2004). Given natural variability, apparent trends in measured data should be treated with caution lest they be revealed as a portion of a long-term cycle (Cohn and Lins, 2005).

Alongside the reality of climate variability is the potential for long term trends due to climate change. Climate science indicates that greenhouse gas emissions are likely to increase global average temperature (Forster et al., 2007), with numerous effects on the hydrologic cycle (Covey et al., 2003; Meehl et al., 2007; Donohue et al., 2010; Trenberth, 2011; McVicar et al., 2012). As mentioned in Section 1.1, reduced precipitation is expected in many regions (Chiew et al., 2009; Gao and Giorgi, 2008; U.S. Global Change Research Program, 2009; Park Williams and Funk, 2011), in contrast to tropical regions which will likely see increases in rainfall (Trenberth, 2011). Increased global temperatures may lead to higher evaporation and transpiration, although this depends on factors including vegetative response to increased carbon dioxide (Keenan et al., 2013) and trends in wind (McVicar et al., 2012). The magnitude of these changes is uncertain and depends upon the models and methods used (Brown et al., 2012; Muerth et al., 2013; Bosshard et al., 2013; Greve et al., 2014), but in some areas there is considerable consensus among models, at least regarding the direction of future changes, as noted above (Chiew et al., 2009; Forzieri et al., 2014; Addor et al., 2014).

Some researchers have argued climate change impacts are already evident in recent weather and climate, although this is difficult to prove conclusively (Hulme, 2014). Greenhouse-induced higher temperatures (Alexander et al., 2006) likely exacerbated recent droughts in some regions including central and west Africa (Niel et al., 2003), south-east Australia (Murphy and Timbal, 2008; Nguyen et al., 2015) and south-west USA (Diffenbaugh et al., 2015). In south-west Australia, recent dry decades have been linked to greenhouse-induced changes in large scale circulation
2.1. **INTRODUCTION**

(Hope et al., 2006). Such evidence underscores the urgent need for robust water resource planning.

2.1.2 **Water resource planning for changing climatic conditions**

Water resource planning is essential to ensure the ongoing security of water supply for domestic, agricultural, industrial and environmental needs. Planning should consider historical behaviour and observations, but also future changes such as to climate conditions and human water consumption. Water resource planners commonly seek to answer questions like:

- In the future, will there be enough water to meet human and environmental needs, and where will this water come from? (Waage and Kaatz, 2011; Gleeson et al., 2012)

- Is current infrastructure appropriate/sufficient/required to store and deliver water in the future? (Bates et al., 2008)

- Are current policies (eg. operating rules, water licencing rules) appropriate for the future? (Bates et al., 2008; van Dijk et al., 2013)

- How best can water be shared between different political priorities? (eg. between different regions, between different usage types (Ostrom et al., 2010; Grafton et al., 2013), between human and environmental needs (Richter et al., 1997; Hughes, 2003; Brauman et al., 2007; Crase et al., 2012))

- How can demands and interactions from different players and sources be managed to minimise risk? (Wilby and Dessai, 2010; Thompson et al., 2013; Aghakouchak et al., 2014; Vogel et al., 2015; Van Loon et al., 2016)

- How certain are the answers to the above questions, what is causing uncertainty, and what does uncertainty mean for decision making? (de Neufville, 2003; Wilby, 2005; Lempert et al., 2006)

To help answer these questions, planners typically employ numerical models. Such models use mathematics to estimate changes with time in various aspects of the
water system. For example, in the context of water resource planning for future climate change, relevant numerical models include: (i) global climate models (GCMs), which estimate precipitation and temperature under future greenhouse emissions scenarios (Wilby and Dessai, 2010); (ii) rainfall-runoff models, which estimate runoff resulting from GCM outputs; and (iii) water systems models, which investigate optimal operation of infrastructure, given projected runoff, and may incorporate adaptation strategies (Maier et al., 2016). The second category, rainfall-runoff models, is the focus of this research. All three model types are used for ‘projection’, which allows managers to answer ‘what-if’ questions. Projection by rainfall-runoff models is referred to here as ‘hydrological projection’.

2.1.3 What is hydrological projection?

Beven and Young (2013) define ‘projection’ as the simulation of the future behaviour of a system given prior assumptions about future input data. An example of projection is the simulation of the future behaviour of the earth system using a GCM, with the prior assumptions contained in greenhouse gas emissions scenario(s). Although some hydrologists use the word ‘prediction’ instead of projection, this word sometimes describes other uses of simulations such as forecasting (Beven and Young, 2013), and is therefore ambiguous.

Hydrologic projection refers to projection applied only to the hydrologic system, usually in a given catchment or region. The inputs to such a system are climatic variables (e.g. rainfall and potential evapotranspiration - PET) and may also include assumptions about other factors (e.g. landuse changes). The key output is runoff, which is calculated using some mathematical description (model) of the river catchment(s). In the context of climate change, climatic inputs are usually derived from the outputs of GCM projections (Chiew et al., 1995, 2009; Bergström et al., 2001; Christensen et al., 2004; Faramarzi et al., 2013; Forzieri et al., 2014).

In summary, hydrological projection is the simulation of the future behaviour of a catchment or catchments, given prior assumptions about future climatic and other variables.
2.1.4 Approaches to hydrological projection

Hydrological projection is often undertaken using numerical models, as mentioned. However, some aspects of hydrological projection can also be undertaken using statistical techniques. This subsection discusses both approaches.

An advantage of statistical approaches is the relative ease with which they can be applied over large areas. Many studies at a continental scale are based on the concept of elasticity, which assumes that runoff changes can be calculated as a simple function (often linear) of rainfall changes (Vogel et al., 1999; Sankarasubramaniam et al., 2001; Chiew, 2006; Fu et al., 2011; Andréassian et al., 2016). However, working in a region with high interannual variability in climate, Saft et al. (2015) demonstrated that simple relationships may not be stationary in time for a given catchment. Furthermore, the utility of elasticity relationships may be limited if the hydroclimate variability within the flow record is relatively low. However, estimation of runoff elasticity can be aided by considering other catchments in the same region via the ‘space for time’ concept (Peel and Blöschl, 2011). By this logic, if projections in a catchment indicate drier future conditions, lessons can be drawn from nearby catchments that are already subject to drier conditions (Singh et al., 2011; Cooper-smith et al., 2014; Patil and Stieglitz, 2015). Existing space for time studies focus on the USA, so future research could apply ‘space for time’ logic to regions in the world with higher historic annual variability, such as Australia (Peel et al., 2001). On a global scale, Budyko (1974) developed a statistical framework for estimating runoff from rainfall and aridity. The framework considers whether catchments are limited by a lack of rainfall or a lack of energy, and it can be adapted for use in hydrologic projection (Zhang et al., 2008). Variants of this framework also consider vegetation type (Zhang et al., 2001; Roderick and Farquhar, 2011; Li et al., 2013).

In contrast to statistical techniques that provide projections based on statistical relationships directly between inputs (rainfall, potential evapotranspiration) and outputs (streamflow), numerical models may also represent the internal processes of this transformation using mathematics (Zhang et al., 2008). For example, catch-
ments store water from previous rainfall events; a wet catchment has more water in storage than the same catchment when dry (McNamara et al., 2011). In between rainfall events, the catchment loses water due to drainage, evaporation and transpiration (Savenije, 2004). Since water takes time to move from place to place, the timing of a flood peak is delayed and a routing algorithm must be used. Numerical models attempt to represent these processes (storage, drainage, routing, etc.) using mathematical functions, whereas statistical techniques generally ignore them by only considering annual timescales (Chiew et al., 1993). Arguably, models that represent such dynamic processes may be more suitable for extrapolation to future changed conditions, so numerical models are sometimes used in preference to statistical methods even for studies of annual dynamics (eg. Arnell, 1992).

Water resource planning often requires shorter timesteps, to capture seasonal variation and event based infrastructure management. Thus, timescale constraints are a key limitation of statistical techniques for hydrologic projection. Water resource studies typically use daily-timestep numerical models (Chiew et al., 1995, 2009; Christensen et al., 2004; Faramarzi et al., 2013; Forzieri et al., 2014), although monthly models are sometimes used instead (Arnell, 1992; Hughes, 1995; Wang et al., 2011; Peel et al., 2015). Despite some evidence that daily models can benefit from process descriptions that account for within-timestep variability (Kandel et al., 2005; Siriwardena et al., 2008; Western et al., 2011), sub-daily models are usually used only by flood studies and as research tools. Given the water resources context of this study (Section 1.1), this thesis uses the phrase ‘rainfall-runoff models’ to mean daily-timestep numerical models unless otherwise stated.

2.1.5 Concepts and definitions in rainfall-runoff modelling

To conclude the introduction to rainfall-runoff modelling, this subsection defines other concepts and terms relevant to modelling and simulation.

In terms of spatial scale, rainfall-runoff models are commonly applied at the scale of individual river catchments, as in this research. However, some projects attempt larger scale modelling (Vaze et al., 2013; Archfield et al., 2015; Bândossy
2.1. INTRODUCTION

The term ‘model structure’ is used in this thesis to mean the underlying mathematical equations of a rainfall-runoff model, and also includes the numerical implementation of these equations. The equations usually have parameters that may vary depending on the case study, allowing the model structure to be ‘fine-tuned’ to a catchment. The term ‘model’ hereafter refers to a model structure with a chosen set of parameters; that is, a model is a parameterised model structure.

It is common to decide parameter values based on the whether the model output matches with observed data, a process called ‘calibration’ (Section 2.3). Deciding parameter values in this way is common because model parameter values often do not correspond with observable quantities, and those that do (e.g. soil water storage) are often not reliable enough for a _priori_ parameter estimation. Calibration is usually done over a period of the historic record, to allow information from different timesteps to inform the chosen parameters. It is good practice to evaluate model performance over one or more historic periods that were not used in the calibration process. This process is often called ‘validation’ in the literature (Oreskes _et al._, 1994), with alternative names including ‘verification’, ‘confirmation’, and ‘evaluation’. In this thesis, the name ‘evaluation’ will be used.

This ends the introduction to rainfall-runoff modelling. Prior to exploring problems with rainfall-runoff models when applied in changing climates (Section 2.2), the study area is first described. This is appropriate because much literature described in Section 2.2 uses examples from the study area.

2.1.6 Introduction to study area

The study area for this research is the temperate regions of southern and eastern Australia. This subsection provides a general introduction to this region, and Figure 2.1 provides a map of the region, showing the spatial distribution and selected characteristics of the 86 adopted study catchments. Description of how the catchments were selected is deferred to Section 4.3.4, and a full catchment list is provided in Appendix A.
Figure 2.1: (a) Map of study catchments and Köppen-Geiger climate types in Australia (after Peel et al., 2007). (b) Mean annual precipitation over the nondry period, for all study catchments. (c) Reduction in long term average precipitation in the driest 7 consecutive years in the record, compared to the remainder of the record. (d) same as (c) but for runoff. The driest 7 years in (c) and (d) are chosen for each catchment on the basis of runoff, not rainfall.

The hydroclimate of the study area is generally mild (Jones et al., 2009). Temperatures rarely drop below freezing, except in isolated mountain pockets in the south-east. Average daily maximum temperatures during summer months are generally 30°C or less. Average annual rainfall is mostly between 600 and 1500 mm/year. Some parts of the study area are subject to dry summers, particularly in the south-west of Australia (Peel et al., 2007). However, in the south-east and east, precipitation is more evenly spread year-round. The study area is influenced by ENSO which causes significant variations in temperature and precipitation on timescales of 4-7 years (Peel et al., 2002). Other cycles are also important, including the IOD, IPO and SAM, and the interaction between these cycles is an important control on climatic conditions (Power et al., 1999; Kiem, 2004).

Owing to Australia’s aridity and climate variability, Australian catchments often have lower runoff coefficients and are subject to relatively high runoff variability on daily and annual timescales, relative to other continents (Croke and Jakeman, 2001; Peel et al., 2001; Baines, 2005). Furthermore, Australian catchments in tem-
2.1. INTRODUCTION

Perate regions are subject to relatively severe runs of precipitation and runoff below the median, compared with non-arid catchments on other continents (Peel et al., 2004, 2005). Such runs can be used to test the ability of rainfall-runoff models in changing climatic conditions. Two prominent examples are the so-called ‘Millennium’ drought in south-east Australia; and the long-term drying of south-west Australia since the 1970s.

The Millennium Drought effected large areas of south-east Australia, both coastal and inland (Verdon-Kidd and Kiem, 2009; Potter et al., 2010; van Dijk et al., 2013). It was particularly unusual for the persistence of the dry conditions, which in some areas last thirteen years (from mid 1997 to mid-late 2010). In terms of rainfall reductions, this drought was comparable to previous droughts in the instrumental record (Potter and Chiew, 2009). However, runoff reductions in many places were unprecedented in the historic record, with one study estimating the return period as 300 years (Potter et al., 2010, cf. Gallant et al., 2011). Rainfall seasonality also altered, partly through reductions in autumn rainfall (Murphy and Timbal, 2008; Potter and Chiew, 2009; van Dijk et al., 2013). Studies using models to investigate causes of streamflow reductions found that changes in rainfall variability and seasonality contributed to low flows (Potter and Chiew, 2011) and likely reductions in groundwater recharge (Larsen and Nicholls, 2012). The Millennium Drought had numerous impacts on Australian society, including cessation of irrigation in some areas causing changes in rural communities, revision of water allocation arrangements to include water trading and provision for environmental flows, and installation of alternative water sources such as desalination in the cities of Melbourne, Sydney and Brisbane (van Dijk et al., 2013; Aghakouchak et al., 2014).

In the south-west of Australia, rainfall, streamflow and groundwater stores have been in decline since the 1970s (Petrone et al., 2010; Hughes et al., 2012). Rain-bearing synoptic troughs have decreased in frequency, while stable high pressure systems have become more common, consistent with the effects of climate change (Hope et al., 2006). In response, streamflow has declined by up to 75 % (Petrone et al., 2010), and catchment average groundwater levels have dropped by up to 6 meters in
upland catchments (Hughes et al., 2012) and up to 4m in the lowlands (Yesertener, 2005), although the latter may also be associated with increased groundwater use. As in the east, municipal water supply for the city of Perth has thus shifted away from surface water, towards both desalination and increased reliance on the declining groundwater resource (Yesertener, 2005).

Climate is projected to get hotter and drier across southern and eastern Australia (Trenberth, 2011; Whetton et al., 2016). Although projections are uncertain and sensitive to assumed greenhouse gas emissions, there is a high degree of agreement among GCMs as to the direction of likely changes, particularly in the south (Chiew et al., 2009). Australia’s drought history and projected future conditions thus provide a vivid example of the need for hydrologic models that can operate reliably in changing climatic conditions.

2.2 Problems with rainfall-runoff modelling in changing climatic conditions

2.2.1 The differential split sample test

To describe problems with rainfall-runoff models in changing climatic conditions, it is first necessary to outline how they are commonly tested. An important assumption of projection studies is that hydrologic models calibrated on historical data are valid to simulate the future, even if future conditions are unlike those in the past (Wagener et al., 2010). While it is not possible to directly test the validity of this assumption, an alternative test is to split up the historic data based on climatic conditions, and calibrate models to a subset of historic climatic variability. Model performance over a different set of climatic conditions then provides an independent test of whether the model can operate under change, a test referred to as the Differential Split Sample Test, DSST (Klemes, 1986; Refsgaard et al., 2014). Klemes (1986) referred to a model that passes a DSST as ‘operationally adequate’, stating that the aim was ‘merely to assess the performance of the model in situations as close as possible to those in which [it is] supposed to be used in practice’. He also extended this logic to
models being translocated in space via the ‘proxy basin test’ (Donnelly-Makowiecki and Moore, 1999; Blöschl et al., 2013). Klemes (1986) presented these tests as a first test of model adequacy, and their fulfillment a ‘necessary, rather than a sufficient, condition for model adequacy’. Nonetheless, rainfall-runoff models often fail the DSST, as discussed below.

2.2. PROBLEMS WITH MODELLING IN CHANGING CLIMATE

2.2.2 Reported differential split sample test results

This subsection focusses on studies using the DSST to test models in changing climatic conditions, although it is noted that hydrologic model performance commonly deteriorates on any period not used in calibration, even if climatic conditions are similar to the calibration period (Vaze et al., 2010). One of the earliest DSST studies was by Refsgaard and Knudsen (1996) who tested a simple conceptual model and two physically based models in three catchments in Zimbabwe. They reported poor simulations during drought, for all three models. Hartmann and Bárdossy (2005) applied a lumped conceptual model to a 2000km$^2$ catchment in Germany, calibrating successively to ‘wet’, ‘dry’, ‘warm’ and ‘cold’ years. They found that models calibrated to the wet years systematically overestimated flow during dry years unless the objective function explicitly included performance measures calculated over longer (annual) timesteps. This indicates that the choice of calibration method may influence rainfall-runoff model performance in changing climatic conditions. Wilby (2005) randomly generated parameter sets to simulate flow in the Thames River, UK, and for each of thirty successive years he identified the best performing parameter set. The sets from wet years were robust and performed relatively well during dry years, but not vice versa. He attributed this to the higher information content of wet years compared to dry years (Vrugt et al., 2002; Wagener, 2003). The findings of Wilby (2005) are generally the exception to the rule, particularly for studies in Australia, which usually conclude that declines in performance are greater going from wet to dry, than vice versa (Vaze et al., 2010; Coron et al., 2012; Li et al., 2012; Silberstein et al., 2013). Vaze et al. (2010) applied four conceptual rainfall-runoff models to 61
catchments in south-east Australia and reported that the decline in DSST performance was generally proportional to the difference in climatic conditions between calibration and evaluation periods, with a greater decline if the change was from wet to dry. Saft et al. (2016a) reported that simulations in DSST evaluation periods were particularly biased in catchments where the annual rainfall-runoff response was shown to be non-stationary (cf. Saft et al., 2015).

Some studies have proposed limits on how much rainfall may change without compromising the performance of calibrated models. Vaze et al. (2010) considered that models in south-east Australia were generally valid provided rainfall was no more than 15% less or 20% more than the calibration period. Singh et al. (Singh et al., 2011) made similar findings, with an acceptable change of 10% drier or 20% wetter for five catchments across the continental USA. In their study of two Irish catchments, Bastola et al. (2011) concluded that changes in precipitation less than 10% would not affect model performance. To limit extrapolation, some studies recommend calibrating models on wet historic periods if a wetter future climate is expected; and dry periods if a drier climate is expected (Li et al., 2012; Broderick et al., 2016). The utility of this approach depends on the hydroclimatic variability in available data, and how far projected climate diverges from this envelope.

2.2.3 Nonstationary parameters

As stated by Merz et al. (2011), “the general idea of parameters in dynamic models is that they represent the stable catchment conditions while the rainfall and other inputs are the time-varying boundary conditions”. In contrast to this paradigm, numerous studies have reported that numerically optimal parameter values change with time when models are calibrated to short historic segments, often changing systematically according to the average climatic conditions in the historic segment. Working in the UK’s Thames basin, Wilby (2005) reported that parameters of daily CATCHMOD models were climate sensitive when the model was calibrated to different two-year periods, particularly for parameters relating to rainfall acceptance at the soil surface. Likewise, Merz et al. (2011) applied daily HBV models to 273 catchment in Austria.
2.2. PROBLEMS WITH MODELLING IN CHANGING CLIMATE

Calibrating to different segments of five years, they found that parameters relating to snow melt and the nonlinearity of runoff generation showed significant correlation with climatic variables such as temperature.

Some studies have utilised the predictability of parameter trends to diagnose structural deficiencies and/or to improve simulations by allowing parameters to change with climatic conditions (Oudin et al., 2006; de Vos et al., 2010; Pathiraja et al., 2016). Other authors have studied the increase in uncertainty in projections resulting from nonstationary parameters (Brigoè et al., 2013). On a deeper level, these results have caused a review of the paradigms (particularly that of stationarity) through which catchments and water systems may be viewed (Clarke, 2007; Milly et al., 2008; Wagener et al., 2010; Ehm et al., 2014; Savinje and Hrachowitz, 2017; Koutsoyiannis, 2011; Koutsoyiannis and Montanari, 2015). Pragmatic interim solutions have included the use of ensembles of model structures to minimise exposure to the structural errors of any one model (Clark et al., 2008; Seiler et al., 2012; Seiler and Anctil, 2015; Broderick et al., 2016; Viney et al., 2009), and trying to limit the degree of extrapolation, as mentioned above. Contrary to the results for daily models, it is noted that studies using monthly models by Niel et al. (2003) and Xu (1999) report time-stable model parameters. Thus, a comparison of the robustness of daily and monthly rainfall-runoff models to changing climatic conditions could be a potential research topic.

In line with the above, rainfall-runoff model structures appear able to operate in a wide variety of conditions, provided parameters can vary. That is, in the hypothetical of five wet years followed by five dry years, the same model structure can generally provide a good fit to both five year periods - just not with the same parameter set. For example, Vaze et al. (2010) calibrated four model structures to the driest ten and the wettest ten year periods on record, for 61 catchments in south-east Australia. They found that the ability to match flows during the calibration period was similar regardless of the period chosen. The key step was then applying the calibrated parameter sets to different climatic regimes, which generally resulted in lower performance.
2.2.4 Characterising deficiencies

It is noted that detailed analyses of simulation weaknesses are rare in DSST studies, which hinders diagnosis of problems (Reusser et al., 2009; Gupta et al., 2009, 2012; Biondi et al., 2012). For example, few studies have reported DSST simulation deficiencies in terms of hydrologic signatures (statistics - cf. Section 2.3.6), which are strongly recommended as a diagnostic tool by Gupta et al. (2009) among others. By way of exception, one hydrologic signature is commonly reported: the mean. Numerous studies have reported overestimation of the mean during dry periods, as mentioned (Refsgaard and Knudsen, 1996; Hartmann and Bárdoossy, 2005; Vaze et al., 2010; Saft et al., 2016a).

To counter this lack of diagnostic information, recent workshops have formulated sets of evaluation criteria (Thirel et al., 2015a) which characterise various aspects of the flow regime such as mean, variability, and low flow behaviour. They also recommend multiple evaluation periods to better understand temporal dynamics (cf. Coron et al., 2012). Given these recommendations, future studies may be more descriptive in reporting DSST results.

Reported deficiencies are not universal across all catchments. Models may work well in some catchments and poorly in others, and literature suggests that it can be difficult to say a priori which model structure will work in which catchment type (Lee et al., 2005). However, the studies by Saft et al. (2015, 2016a, 2016b) provide some understanding of which catchment properties tended to be associated with model failure for the case of the Millennium Drought. Analysis of shifts in the annual rainfall-runoff relationship during this drought revealed that flatter and less forested catchments (Saft et al., 2015) with thin soils and variable year-to-year groundwater storage (Saft et al., 2016b; Van Loon and Laaha, 2015) were more likely to shift away from historic rainfall-runoff relationships. Furthermore, catchments of greater aridity prior to the drought are more susceptible to subsequent reductions in rainfall (Arnell, 1992). Saft et al. (2016a) confirmed that the shifts in annual relationship were highly correlated with poor rainfall-runoff model performance.
2.2. PROBLEMS WITH MODELLING IN CHANGING CLIMATE

2.2.5 Possible causes for deficiencies

What is causing the decline in performance of rainfall-runoff models when climatic conditions change? Some authors have cited the need to improve model structures (Merz et al., 2011; Rannacharn, 2012; Silberstein et al., 2013; Potter et al., 2013; Hughes et al., 2013). However, the issue of characterising model performance is complicated (Reusser et al., 2009; Bennett et al., 2013) as is the attribution of causes of poor model performance, so there may be other factors involved (Shoaib et al., 2016, 2018). Coron et al. (2014) conclude their study by listing the following potential causes (p741):

a an ineffective model structure;

b an inappropriate calibration strategy;

c temporal changes in input errors;

d temporal changes in the catchments’ natural functioning; and/or

e temporal changes in anthropogenic impact.

The remainder of this literature review examines some of these items in more detail. Section 2.3 discusses calibration methods, while Section 2.4 discusses the three related problems of model structural inadequacy, anthropogenic impacts, and changes in natural functioning. Although data problems (c above) are a possible explanation for a given case study, it seems unlikely that data problems are causing so many model failures across such geographically diverse regions, so they are not the focus here (although they are discussed at other points in the thesis - eg. Section 4.5.3 and Section 7.4.9).

2.2.6 Summary

In summary, hydrologists often use the DSST as a first test of whether rainfall-runoff models can simulate under changing climatic conditions. Numerous studies from different regions of the world attest that models often fail the DSST, producing biased simulations after a change in climate. A related problem is parameter
non-stationarity, whereby numerically optimum parameters change according to the climatic conditions in the calibration data. The discussion now turns to how we define ‘numerically optimum’, which requires a review of common calibration methods.

2.3 Calibration of daily rainfall-runoff models

Sorooshian and Gupta (1995) define calibration as the process of “selecting values of the model parameters so that the model closely simulates the behaviour of the study site” (p23). Arising from this definition are various questions. Firstly, what is meant by ‘closely’? Secondly, how should ‘close’ parameter sets be identified? Thirdly, what should be done when simulations from many different parameter sets are nearly as ‘close’ as one another? Lastly, how do the answers impact simulations in changing climates? These questions are explored in this section.

2.3.1 Manual calibration

When simulation models were first developed, modellers would arrive at suitable parameter values via manual trial and error. The definition of ‘closeness’ was based on modellers judgement of whether the model was replicating dominant processes, typically based on graphical comparisons. Manual calibration was necessary because early computer power was insufficient to support automated procedures. However, manual calibration has other benefits not related to computational efficiency (Gupta et al., 1998; Boyle et al., 2000; Wagener, 2003). The human brain can consider many criteria simultaneously and with a subtlety which can be difficult to automate (Gupta et al., 1998). Human pattern recognition is highly developed, even to the point where some automated procedures seek to mimic the steps humans take intuitively (Friedman and Tukey, 1974). Manual calibration can account for potential errors in input data where automated procedures would treat all values as equally valid. However, manual calibration can be exceedingly labour intensive (Boyle et al., 2000; Smith et al., 2003) and subjective because each modeller may have their own idea of ‘closeness’ (Ibbitt and O’Donnell, 1971). These factors, in addition to historical
perceptions that automated processes would prove superior (Wagener, 2003), has led to manual calibration becoming relatively rare in practice (but cf. Boyle et al., 2000; Smith et al., 2003).

### 2.3.2 Automatic calibration

As computing power increased, more model runs became possible, facilitating automatic calibration. Automatic calibration requires that the ‘closeness’ of model simulations - previously evaluated using human judgement - be formulated into some mathematical function (objective function). It also requires an algorithm for searching the parameter space to identify the parameter set with the best possible objective function score, a process known as optimisation. Initially, only optimisation to a single objective was available, but methods for considering multiple objectives were developed later (Section 2.3.6). The advantages of this approach were that it was relatively objective (provided objective functions could be agreed) and theoretically repeatable, which aids defensibility (Hutton et al., 2016). Furthermore, it reduced reliance on highly skilled individuals trained in the use (and idiosyncrasies) of a given model structure (Boyle et al., 2000).

As experimentation progressed, it became clear that automatic calibration is fraught with potential difficulties. As described in early studies such as Johnston and Pilgrim (1973), some parameters did not appear to influence the result, so that any adopted value would be arbitrary (called non-identifiability), while others were subject to large discontinuities, so that small changes in parameter value resulted in large changes in objective function value (Clark and Katetski, 2010a). Optimisation algorithms would sometimes become ‘trapped’ in local optima (eg. Duan et al., 1992), compromising repeatability and giving a false negative impression of model capabilities (phrased as ‘mis-calibration’ by Andréassian et al., 2012; Duan et al., 1992). Also, often many parameter sets were very close to optimal (equifinality: Beven, 2006), sometimes from separate parts of the parameter space.

These problems may help to explain why early implementations were criticised based on unreasonable parameter values and poor simulations (Boyle et al., 2000).
Nonetheless, helped by advances to lessen (or at least understand) some of these problems (Duan et al., 1992; Boyle et al., 2000; Sobol, 2001; Kavetski and Kuczera, 2007; Clark and Kavetski, 2010a; Kavetski and Clark, 2010) automatic calibration has become common practice (Wagener, 2003). Many of the problems were shown to be exacerbated by overparameterisation, which means that models have more parameters than can be justified by the context and training data (Grayson et al., 1992; Jakeman and Hornberger, 1993; Grayson and Blöschl, 2001; Schoups et al., 2008). Thus, the development of parsimonious models (Boyle et al., 2001; Perrin et al., 2001; Chiew et al., 2002), together with optimisation algorithms more skilled at working in highly irregular response surfaces, has progressed automated calibration. Today, many optimisation algorithms are available (Gupta and Sorooshian, 1985; Duan et al., 1992; Hansen et al., 2003; Arsenault et al., 2014), and many different objective functions have been trialled, as discussed below.

2.3.3 Choosing an objective function

Many authors have argued for a strictly statistical approach to objective function selection. For example, classical statistics would suggest that the parameter set that is statistically most likely (under certain assumptions) will be the one that minimises the sum of the square of model residuals. In this scheme, the difference between the simulated and observed runoff (residual) on each timestep is squared, and then the sum of these squares over the entire modelling period is the objective function to be minimised. Objective functions based on this logic are referred to as ‘least-squares’ metrics, and include the Nash Sutcliffe Efficiency (NSE, Nash and Sutcliffe, 1970)) and the Root Mean Square Error (RMSE, eg. Gupta et al., 1998).

Despite their popularity, least squares metrics have several drawbacks. They are often applied with little consideration of the assumptions of the underlying statistical theory (Legates and McCabe, 1999; Kruze and Boyle, 2005; Criss and Winston, 2008). For example, assumptions that residuals have zero autocorrelation and that the variance of error does not change with flow magnitude (homooscedasticity) are rarely met in hydrological practice (Sorooshian and Dracup, 1980; Kuczera, 1983).
2.3. CALIBRATION OF DAILY RAINFALL-RUNOFF MODELS

This skews selection of parameter sets, and literature suggests that least-squares methods tend to emphasise performance during high flow periods (Gan et al., 1997; Legates and McCabe, 1999) at the expense of low or medium flow periods. Furthermore, Gupta et al. (2009) showed that least squares methods can be decomposed into components (cf. Murphy, 1988) and problems can arise from interactions between components. For example, high NSE scores may mask unacceptably high bias and low variance relative to the observed variance (see also Tian et al., 2016). Common use of NSE values has further been criticised because the NSE formulation requires the definition of a reference model, and most users simply adopt the long-term mean as a reference model, without considering alternative models which, as argued by Seibert (2001), may be more appropriate in highly seasonal catchments.

Given these drawbacks, alternatives to least squares metrics have been proposed. Some authors apply transformations to flow values prior to least-squares-type calculations, in order to stabilise the variance (Engeland et al., 2005); such transformations also place emphasis on different types of hydrological behaviour (Krause and Boyle, 2005). For example, more emphasis on low- and mid-flows was achieved by Chiew et al. (1993) via transforming flows by the fifth root (raising to the power of 0.2) prior to least squares calculations, and by de Vos et al. (2010) by taking the logarithm of flows. Other examples include metrics that emphasise a match with low flows via taking the inverse of flows prior to least squares calculations (Pushpalatha et al., 2012) or a more balanced consideration of ‘average’ model performance through absolute-error, rather than squared-error, approaches (Willmott, 1982; Willmott and Matsuura, 2005).

2.3.4 Consideration of uncertainty

Irrespective of whether manual calibration or automatic optimisation are chosen, both methods choose a single parameter set at the expense of all the others. Many authors have argued that this faith in a single parameter set is misguided (Uhlenbrook et al., 1999; Wagener et al., 2001), given the multiple different sources of uncertainty in hydrology.
Firstly, all measurements of environmental variables are subject to uncertainty (Kuczera et al., 2010). Rain and flow estimates are subject to errors even when best practice measurement is used; and often have significant epistemic errors due to instrument neglect, malpractice, malfunction, or inappropriate post processing (Nešpor and Sevruk, 1999; Andréassian et al., 2001; Michelson, 2004; Sevruk et al., 2009; di Baldassarre and Montanari, 2009; McMillan et al., 2010; Cozon et al., 2015). Further, environmental data are rarely available at exactly the location required, leading to further epistemic error due to interpolation or extrapolation, particularly when local data are scarce (Obled et al., 1994; Das et al., 2008; Kuczera et al., 2010). Most rainfall-runoff models also use potential evapotranspiration (PET) data, which depend strongly on the particular mathematical formulation adopted (Oudin et al., 2005; Donohue et al., 2010; McMahon et al., 2013; Seiller and Anctil, 2015) and are often the subject of debate in the literature due to inappropriate simplification and conceptualisation (Savenije, 2004; McVicar et al., 2012; Milly and Dunne, 2016).

Parameter identification can be highly sensitive to errors in forcing data and training data (Andréassian et al., 2001; Oudin et al., 2006). Boughton (2006) states that the “results from rainfall-runoff modelling are more dependent on the quality of the input data than on the model [structure]”. Berthet et al. (2010) showed how a small number of large events can have a major impact on the objective function value. A different set of errors during these events may lead to a different parameter set being mathematically optimum, sometimes from a distant part of the parameter space, with different projections if applied in changed climatic conditions. Thus, uncertainty in inputs affects parameter identifiability, although some studies have proposed methods to separate the two, as discussed in Section 2.3.5. There remains inherent structural uncertainty due to the simplification of the real system by the model (Butts et al., 2004; Shin et al., 2015), but it is commonly augmented by input and output uncertainty, when these are not explicitly considered in the calibration process.

Uncertainty in outputs is related to the information content of the calibration data. While even a handful of runoff measurements can be beneficial to parame-
ter estimation (Seibert and Beven, 2009), a key benefit of longer timeseries is the potential for capturing different modes of hydrologic behaviour. If a model represents an important process that is not active during the calibration period, then the parameter(s) that govern that process will be poorly identified. Although some literature suggests that such parameters should be fixed (van Werkhoven et al., 2009), leaving such parameters free reflects the reality that the process remains ill-defined; if the process becomes active and important during evaluation, fixing such parameters may underestimate predictive uncertainty if ensemble methods are used (Section 2.3.5, Reichert and Omlin, 1997; Wagener, 2003).

An extra source of uncertainty is the subjective modelling decisions made by the modeller, as demonstrated by the comparison of five consulting studies given the same brief by Refsgaard et al. (2006). The final results of each study bore no resemblance to the others, mainly due to methodological choices. Even in cases where methods are purported to be the same, subjective (often unreported) choices between modellers can lead to difficulties reproducing results, which has led to calls for standardised approaches to modelling and more open publication of code and data (Hutton et al., 2016; Koutsoyiannis et al., 2016).

### 2.3.5 Ensemble methods

As mentioned, there may be other parameter sets that give almost as good performance as the numerically optimum set (equifinality - Beven, 2006). Rather than ignoring this issue, numerous authors argue that these parameter sets must be accepted as plausible alternative model hypotheses (eg. Beven, 2012). They thus propose consideration of multiple parameter sets, called an ensemble. It is noted that this logic can be extended to other plausible model structures also (Perrin et al., 2001; Clark et al., 2011, 2016).

A key advantage of ensemble methods is that they can facilitate a statistically rigorous approach to estimation of output uncertainty, particularly when coupled with error models that correct certain tendencies usually present in inputs and outputs of rainfall-runoff simulations, such as transformations to stabilise the variance
of errors, autocorrelation models, and event-based rain error models (Box and Cox, 1964; Sorooshian and Dracup, 1980; Kuczera, 1983; Engelund et al., 2005; Kuczera et al., 2006). These models contain additional parameters (Schaink and Vrugt, 2010), particularly when correcting rainfall errors (Kuczera et al., 2006; Kavetski et al., 2006a,b). Authors such as Renard et al. (2010; 2011) and Thyer et al. (2009) argue that these methods allow better quantification and characterisation of error sources and parameter uncertainty. However, these methods are relatively complex and users need well developed statistical knowledge. In addition, the theoretical underpinning of some aspects has been debated (Beven, 2011).

In part due to these disadvantages, many hydrologists advocate the use of simpler ‘informal’ statistical methods that choose ensembles from random samples of parameter sets, based directly on objective function score (eg. GLUE - Beven and Binley, 1992) or on whether simulations lie within pragmatic bounds of plausibility considering possible data errors (‘Limits of Acceptability’ - Beven, 2006; Blazkova and Beven, 2009). Such methods are more feasible to apply widely, but are subject to criticism for the lack of explicit separation of error sources (Beven and Freer, 2001a), the need for subjective decisions such as the selection of thresholds, and perceived inconsistencies with statistical theory (Mantovan and Todini, 2006; Beven et al., 2008), despite similar modelling results in some case studies (Vrugt et al., 2009).

Ensemble methods can be used to inform model averaging techniques, where the results of multiple models are combined for potentially improved projections (eg. Oudin et al., 2006), and ‘mixture of experts’ approaches that allow the degree of influence of different models to vary as conditions change (eg. Marshall et al., 2007). In the latter approach, a set of pre-calibrated models are run with the same forcing data, and models are weighted for each timestep in a way that reflects the success of the models over the calibration data. For example, a model may have performed better when the catchment was wet but not dry, and thus the projections from this model are given a relatively greater weight at wet times. At its extreme, this approach allows ‘switching’ between models at different times and under different conditions.
2.3. CALIBRATION OF DAILY RAINFALL-RUNOFF MODELS

Difficulties in this approach lie partly in predicting when to switch between models under assumed future conditions, which could potentially be informed by data assimilation techniques (Pathiraja et al., 2016). Also, some hydrologists object to the dishonouring of water balance principles when switching from one model to the other.

In practice, even the simpler methods discussed above (eg. GLUE) might be too time consuming to implement in studies that consider large samples of catchments and/or model structures; or applications where the rainfall-runoff simulations are themselves input into complicated analyses, eg. water resources planning tools (Vaze et al., 2011), stochastic decision support tools or ecohydrological investigations (Ramchurn, 2012).

2.3.6 Multi-objective calibration and hydrologic signatures

Most of the automatic calibration methods discussed above use a single objective in calibration. This inherently causes a loss of information, as multiple nuanced aspects of simulation ‘closeness’ must be collapsed into a single number (Wagener, 2003). Considering multiple objectives in calibration may be a better approach, for two main reasons.

Firstly, multi-objective approaches may extract more relevant information out of the calibration flow data (Gupta et al., 1998; Wagener et al., 2009). Gupta et al. (1998) recommended that objectives should be relatively independent from one another, characterising different aspects of the flow regime. Multi-objective approaches allow objectives that are specific to the problem at hand, such as low flow performance in ecological studies (Pushpalatha et al., 2012), to be considered alongside more traditional metrics. Some studies use time-varying calibration techniques, that allow the separate characterisation of errors over different time periods (eg. seasons) within the calibration period (Kim and Lee, 2014; Choi and Beven, 2007). Smith et al. (2008) used a particle filtering approach that ‘reveals the parameter distribution needed at each time to reproduce observed data’, using similar concepts to those in the Dynamic Identifiability Analysis (Dyna) of Wagener (2003). Reusser et al.
(2009) extended this time-variable approach to consideration of multiple metrics in each time segment, using self-organising maps to categorise common error types. Gharari et al. (2013) explicitly defined trade-offs between different objectives in different periods, using a Pareto approach (see below). All these methods are likely to extract more information from observed flow data, compared to using a single objective.

Secondly, multi-objective approaches can facilitate calibration on more than one data type. Examples include water quality information (Kuczera and Mroczkowski, 1998; Mroczkowski et al., 1997), groundwater levels (Lamb et al., 1998; Seibert et al., 1997; Seibert, 2000) and soil moisture measurements (Kalma et al., 1995). In addition, various types of ‘soft’ data may be incorporated, including fuzzy rules about simulated behaviour, expert judgement, and sporadic ‘snapshot’ data, eg. from remote sensing (Seibert and McDonnell, 2002; Winsemius et al., 2009; Gharari et al., 2014). Although multi-response data do not always reduce parameter uncertainty (Kuczera and Mroczkowski, 1998), Mroczkowski et al. (1997) emphasised their value for evaluating catchment models simulating shifts in hydrologic regime. Furthermore, multivariate data may be helpful to make the model structure itself more realistic: "[the ]scrutiny of multivariate data ... not only exposes major model deficiencies, but may be indispensable for improving model realism" (Clark and Kavetski, 2010a). Thus, improved calibration is only one of many reasons for additional data types in the calibration process, assuming such data are indeed available.

Many different methods of multi-objective calibration are available (Efstratiadis and Koutsoyiannis, 2010). Some authors simply optimise a single ‘meta’ objective based on a weighted combination of multiple objectives (Zhang et al., 2008), but this may mask hidden trade-offs among the competing components. trade-offs can be explicitly explored by identifying the set of parameter sets called the Pareto Front (Pareto et al., 1972; Ritzel et al., 1994; Cieniawski et al., 1995; Yapo et al., 1998; Madsen, 2003; Vrugt and Robinson, 2007; Wöhling et al., 2013). All parameter sets on the Pareto Front have the property that the score in one objective cannot be increased without decreasing the score in another objective. Alternative approaches
2.3. **CALIBRATION OF DAILY RAINFALL-RUNOFF MODELS**

use rejectionist frameworks, where all objectives must be above selected thresholds (Pechlivanidis et al., 2014). The ‘limits of acceptability’ approach mentioned above (Beven, 2006) is a rejectionist framework where each day constitutes its own objective (cf. Gupta et al., 1998).

Other multi-objective methods use hydrologic signatures, which are statistics describing the flow timeseries. Early studies used hydrologic signatures to characterise elements of the flow regime relevant to ecosystem health (Clausen and Biggs, 2000; Olden and Poff, 2003), and later, prediction in ungauged basins (Yadav et al., 2007; Bárdossy, 2007; Wagener and Montanari, 2011). Model calibration studies have demonstrated that signatures can help to: extract calibration data on different timescales (Shamir et al., 2005a); restrict the behavioural parameter space, aiding parameter identifiability (Yilmaz et al., 2008; Bulygina et al., 2009, 2011); assess the realism and consistency of model structures both prior to calibration and during split sample testing (Parkin et al., 1996; Vogel and Sankarasubramanian, 2003; Wagener et al., 2007; Bai et al., 2009; Herbst et al., 2009; Enser et al., 2013; Shafii and Tolson, 2015); contribute to model-based assessments of hydrologic stationarity (Sadegh et al., 2015); facilitate multi-site calibration (Castiglioni et al., 2010); and enable model calibration focussed on hydrologically meaningful metrics such as the flow duration curve (Winsemius et al., 2009; Westerberg et al., 2011; Pfannerstill et al., 2014). Various methods of signature-based model calibration exist, some of which incorporate signatures into frameworks that include least-squares measures or limits of acceptability (Shamir et al., 2005b; Winsemius et al., 2009), while others use signatures exclusively (Diggle and Gratto, 1984; Marjoram et al., 2003; Vrugt and Sadegh, 2013; Sadegh et al., 2015). According to Vrugt and Sadegh (2013), these latter signature-only methods can be more defensible statistically than commonly used calibration methods. A possible drawback is subjectivity, because the modeller must choose which signatures to calibrate against. However, signatures can be chosen to minimise sensitivity to measurement error (Vrugt and Sadegh, 2013; Westerberg and McMillan, 2015; Westerberg et al., 2016), bearing in mind that a given signature may be relatively more useful for calibration in some catchment types (eg. for karst
Landscapes see Coxon et al., 2014; for glacial catchments see Koch et al., 2014). Hydrological signatures have proved flexible and an insightful addition to many model calibration studies.

In summary, multi-objective methods allow model calibration to incorporate more information, either by extracting more information from a single response series, or facilitating multi-response calibration. Multi-objective tests have the potential to more stringently test a hydrological model than single objective optimisation (Kollat et al., 2011), and failure of such tests, far from being a drawback, is likely to be instructive for model improvement (Kirchner, 2006).

2.3.7 Calibration methods contributing to existing literature consensus

Having considered many aspects of model calibration, this discussion now returns to the context of changing climates. Given the aforementioned literature consensus that rainfall-runoff models are unreliable in changing climates (Section 2.2.2), this subsection considers two questions:

a Which calibration methods were used to establish this consensus?

b Are better calibration methods available?

The answer to the first question is that almost all studies reviewed in Section 2.2.2 used variants of least-squares objective functions, mostly in the context of single objective optimisation. Common choices were the NSE (Wilby, 2005; Li et al., 2012) and closely related KGE (Coron et al., 2014), sometimes subject to a penalty or limit on bias (Vaze et al., 2010; Silberstein et al., 2013; Sæft et al., 2016a). Only one study (Coron et al., 2012) applied a transform prior to least squares calculation, in this case the square root, applied with a bias penalty. Most studies used single objective optimisation, and even studies that worked with random samples made conclusions with reference to the single highest parameter set (Wilby, 2005) or a small subsample with high performance (e.g. top 0.01% in Li et al., 2012), which is likely to give similar results to single objective optimisation. The only exception to all the above points was Refsgaard and Knudsen (1996) who used manual calibration.
2.3. CALIBRATION OF DAILY RAINFALL-RUNOFF MODELS

Thus, the literature consensus is generally based on studies using single objective optimisation to least squares metrics (or similar). With respect to timestep, all of the above studies calculated objective functions on a daily timestep and used daily models.

Are better calibration methods available? A review of the literature suggests multiple avenues for potential improvement. As noted, different objective functions emphasise different aspects of flow behaviour, and least squares metrics tend to focus on matching high flow behaviour. However, high flow behaviour may have limited relevance to projections subject to drying climate. The flow response during low flow seasons or dry years might contain valuable information for parameter estimation, and yet be largely ignored in the calculation of least-squares metrics. Thus, objective functions that emphasise different flow behaviours may be helpful, for example, through transforms (Krause and Boyle, 2005; Oudin et al., 2006; Pushpalatha et al., 2012; de Vos et al., 2010), or considering multiple objectives such as the matching of selected hydrologic signatures (Winsemius et al., 2009).

Likewise, methods that focus on a variety of timescales may provide an improved calibration, as suggested by the following studies. As mentioned, Hartmann and Bárdossy (2005) formulated numerous objective functions based on least squares calculations at different timesteps (eg. daily, annual, decadal). Methods that combined both annual and daily objective functions into a single ‘meta-objective’ were shown to reduce the error in DSST annual flows from 30% to 10%. Likewise, Shamir et al. (2005a) applied multi-timescale logic but based their analysis on hydrologic signatures defined on daily, monthly and annual and multi-annual temporal scales. The result was an ensemble of parameter sets that performed well on all timescales considered; the identifiability of parameters in the Sacramento model was also improved.

A further avenue is the statistical concept of data depth, which Bárdossy and Singh (2008) introduced to hydrological modelling. A parameter set has greater depth if it is located closer to the centre of a cloud of well performing sets (Tukey, 1975). Using data from Germany, Bárdossy and Singh (2008) found that parameter
sets with greater data depth were less sensitive to random errors in input data and thus more robust in DSSTs. The changes tested were up to 10% and up to 25% for rainfall and streamflow, respectively.

Although the literature suggests many alternative objective functions, there are few studies that systematically test these functions in a changing climate. Likewise, despite many alternatives to single objective optimisation, there are few studies that compare more than one method using DSSTs. This is a clear research need, given recent evidence (Seiller et al., 2017) of the difference that the calibration method makes to diagnosis of climate change impacts on water resources.

Lastly, ensemble methods may be useful for diagnosing poor model performance. As noted earlier, Vaze et al. (2010) reported that model structures are often able to match dry periods and wet periods - just not with the same parameter set. However, if parameter sets other than the numerically optimum set are considered, it might be possible to identify parameter sets with sufficient performance in both dry and wet. Testing for these parameter sets is a research gap. Existing research has applied the idea of robust parameter sets across space (Andréassian et al., 2014) but not time. Such robust parameters could be considered closer to the hydrological optimum of (Andréassian et al., 2012), as opposed to the numerical optimum revealed by single objective optimisation. If such parameter sets were found to exist in many case studies, the poor model performance reported in 2.2.2 could potentially be due to poor calibration methods rather than insufficient model structures.

### 2.3.8 Summary of findings for review of calibration methods

In summary, calibration methods are an important aspect of hydrological projections. Although automated calibration methods have the potential to make calibration more objective and repeatable than previously used manual techniques, they have numerous difficulties. It is unclear which objective function is most suitable for changing climates, and commonly used least squares metrics may focus on hydrologic behaviours of limited relevance to drying climates and be unduly influenced by measurement errors during high flow. Flow transformation and consideration of
2.4. IMPROVING RAINFALL-RUNOFF MODEL STRUCTURES

multiple objectives, including hydrologic signatures, may provide more robust simulations. Reliance on single ‘optimal’ parameter sets may provide misleading results, and diagnosis of problems may benefit from ensemble methods. Based on this review, there appear to be many possibilities for improving rainfall-runoff model calibration for changing climates.

2.4 Improving rainfall-runoff model structures

Notwithstanding the importance of parameterisation, the realism of model structures remains an important issue (Bahremand, 2016). It is likely that at least some of the problems noted in Section 2.2.2 are due to model structures themselves, and this is the topic of this subsection. The subsection begins by providing general context for rainfall-runoff modelling, including a discussion of physically based and conceptual models (Section 2.4.1), followed by a review of possible methods to infer structural deficiencies from simulations (Section 2.4.2). Finally, different hypotheses of deficiency are discussed (Section 2.4.3).

2.4.1 Overview of physically based and conceptual modelling

From the earliest stages of hydrological modelling, some authors have advocated for simulation models based on physical laws (eg. Freeze and Harlan, 1969). Initially, it was thought that satisfactory (often laboratory-based) mathematical descriptions of physical laws were available for most components of hydrological response, and the arrival of digital computers would allow these laws to be linked together to simulate runoff response at catchment scale. Indeed, many models today are partially based on this ideal (eg. Abbott et al., 1986; Sulis et al., 2010; Kollet et al., 2010). In theory, model parameters in such models could be directly measured, or based on pre-existing databases compiled for different characteristics and conditions (eg. SWAT - Arnold et al., 1998).

Nearly fifty years later, many difficulties impede physically based hydrological modelling (cf. Beven, 1989; Grayson et al., 1992). For example, measuring soil
CHAPTER 2. LITERATURE REVIEW

characteristics, hydrologic states and fluxes remains difficult to do in-situ (Beven and Binley, 1992; Beven and Germann, 2013; Vereecken et al., 2015) particularly at catchment scale. It is important to understand which hydrologic processes are dominant at which scales (Klemeš, 1983; Beven, 1995, 2006; Blöschl and Sivapalan, 1995; Blöschl, 2001; Western et al., 1999, 2002; Harman and Sivapalan, 2009) and in which hydroclimatic regimes and geologies (Dunne, 1983). Many equations from Freeze and Harlan (1969) were derived in the laboratory, but more recent authors have stated that process-based equations at hillslope scale should be expected to look different to lab-scale equations (Reggiani et al., 2000; Weiler and McDonnell, 2004; Kirchner, 2006; Harman and Sivapalan, 2009). Hydrological processes are highly non-linear (Beven, 2001, 2006; Blöschl and Zehe, 2005) so it is difficult to select representative parameters for spatially heterogeneous catchments (Kling and Gupta, 2009). Catchments may also gain or lose water via sub-surface pathways (Le Moine et al., 2007) and the high uncertainty in environmental data (Section 2.3.4) makes it difficult to close the water balance with confidence. Some processes are rarely well represented, such as the high amount of ‘old’ water in stormflow (Martinec, 1975; McDonnell et al., 1990; Kirchner et al., 2000, 2001; Kirchner, 2003; McGuire and McDonnell, 2006; McDonnell and Beven, 2014). Physically based models have thus been categorised as ‘process weak’ (McDonnell et al., 2007), dependant on data that may not be available in practice (Grayson et al., 1992), and computationally expensive, which may limit calibration options (Reed et al., 2004; Razavi and Tolson, 2013). Progress continues on this class of model, despite these difficulties (Paniconi and Putti, 2015).

Process representation aside, the issue of parameterisation of highly parameterised physically based models remains. In practice, many parameters are adjusted to increase model performance (eg. Reed et al., 2004), leading to a highly dimensional calibration problem. In contrast, the information in a flow series is finite (Daw and Finney, 2003; Weijs et al., 2013) and can only be used to uniquely identify a limited number of parameters (Beven, 1989; Jakeman and Hornberger, 1993; Ye et al., 1997). Thus, some studies have opted for ‘top-down’ approaches, where model com-
plexity is added only if warranted by the data itself (eg. Jothityangkoon et al., 2001; Sivapalan et al., 2003; Chirico et al., 2003; Farmer et al., 2003; Young, 2003; Fenicia et al., 2008, 2011, 2014; Zhang et al., 2008; Smith et al., 2013; van Esse et al., 2013; Gharari et al., 2014; Willems et al., 2014; Markstrom et al., 2016). In such studies, the form of the equations, not just their parameters, may be dictated by the data (eg. Lamb and Beven, 1997). These exercises are often site-specific and sometimes result in remarkably parsimonious models (eg. Kirchner, 2009), but careful judgement is required to include all processes that may be important across all conditions where the model might be applied in future, even if such processes were not important in the training data (Reichert and Omlin, 1997).

Alongside these site-specific studies, the derivation of more general models has remained an important goal. Despite the difficulties of ‘uniqueness of place’ (Beven, 2000), numerous studies have developed simple model structures that can simulate flow in many catchments and contexts (eg. Wood et al., 1992; Linström et al., 1997; Boyle et al., 2001; Chiew et al., 2002; Perrin et al., 2003). These so-called ‘conceptual’ models may represent the concepts of physical processes without reference to physically based equations. Due to the constraints of early computing, the earliest hydrologic simulations models were of this type (eg. Crawford and Linsley, 1966; Burnash et al., 1973; Porter and McMahon, 1976). Such models have fewer data requirements and are less time consuming to set up and run, so they continue to be relatively popular in practice, particularly for studies covering many catchments. Some recent model developments have focussed on parsimony (minimising the number of parameters, Perrin et al., 2001, 2003; Chiew et al., 2002), which may aid problems of equifinality and parameter identifiability. Conceptual models often provide similar or better DSST performance than complex physically-based models (Ræsgaard and Knudsen, 1996), although are sometimes unable to simultaneously match measurements relating to state variables (eg. soil moisture, Alley, 1984; Chiew et al., 1993). Indeed, conceptual model states may not correspond clearly with measurable environment variables, meaning that comparison is fraught with challenges of interpretation (Reægaard, 1997). Conceptual models are often spatially lumped,
thus losing information on internal spatial variability of processes, although some conceptual structures are formulated with this in mind (e.g. Beven and Kirkby, 1979; Wood et al., 1992; Gan et al., 1997; Moore, 2007).

Model application (whether physically based or conceptual) has also responded to trends in data availability. Early studies were often catchments specific, with data estimated from a small number of gauged locations. The increasing availability of spatially distributed datasets (e.g. elevation, climate data) has enabled modelling schemes over large continuous areas or even whole countries (Covey et al., 2003; Vaze et al., 2013; Archfield et al., 2015; Faticchi et al., 2016; Beck et al., 2016). Even in cases of high spatial discretisation, parameterisation is often lumped to the spatial scale of available streamflow data (Muleta and Nicklow, 2005).

In summary, while some studies have demonstrated physically-based hydrological modelling, this approach presents many difficulties, including measuring and representing highly non-linear natural processes occurring on a variety of scales. Many site specific studies have adopted ‘top-down’ approaches, whereby model complexity is added only if warranted by the data. A popular alternative is generic conceptual model structures, developed to be applicable over a wide variety of catchments. The increasing availability of spatial datasets has changed the way all these models are applied.

### 2.4.2 Inferring structural deficiencies from simulation results

Can model structural deficiencies be inferred from simulation results? As noted, DSST studies to date tend to report only on global performance metrics and bias, and inferring deficiencies from this scant information is difficult (Martinez and Gupta, 2010). However, authors such as Gupta et al. (2008) have advocated for a clearer diagnostic approach to model evaluation based on flow statistics (hydrologic signatures) chosen to be as ‘theoretically relevant’ as possible. In this view, comparing the degree of replication across multiple hydrologic signatures gives insight into which aspects of the flow regime are poorly matched, insight not possible with global performance scores. Gupta et al. (2008) then suggest that model improvement can proceed
by deducing (e.g. Sobol, 2001) which component of a model structure relates to a signature shown to be poorly matched, and focussing on that component. de Vos et al. (2010) split historic data into temporal clusters based on hydrologic similarity, calibrating the same hydrologic model structure to each cluster in turn. Assuming that “deficiencies of the model structure cause the model parameters to vary ... to compensate for the effects of the model structural error” (p2841), their method aimed to infer how best to improve the model structure, by observing which parameters changed. Numerous other authors have demonstrated methods for analysing how parameters change to fit sliding windows of time, on short (Wagener, 2003) or long (Merz et al., 2011) timescales. Westra et al. (2014) presented an experimental framework where one or more parameters change value with time based on selected co-variates (Julian day, recent climatic conditions, or a linear trend). If such a change leads to a significant increase in performance, Westra et al.’s 2014 method infers that problems exist with the corresponding model component. This study used only one model and catchment; in the present context, more catchments are potentially required to advance evidence-based insight into structural deficiencies (Gupta et al., 2014; Newman et al., 2015). Nonetheless, these studies indicate simulation-based inference of structural deficiencies is possible. Structural deficiencies can also be identified based on physical and/or theoretical grounds, some examples of which are given in the next subsection.

2.4.3 Possible structural deficiencies

To complete the literature review, possible deficiencies in model structures are discussed below. Note that this subsection focuses on conceptual rainfall-runoff models. This is because (1) hydrological projections are most commonly provided by conceptual models (Section 1.1); and (2) studies that contributed to the literature consensus that models are unreliable in changing climates (Section 2.2.2) predominantly used conceptual models.
CHAPTER 2. LITERATURE REVIEW

Implications of model bias for ET

As noted, hydrological models tend to overestimate flows when evaluated over droughts. This implies underestimation of ET, because the three largest water balance components are typically precipitation, runoff and ET. Thus, structural improvement efforts might focus on model components that govern ET - generally, the soil moisture accounting portions, and (if present) interception stores. This could also include a review of where ET is sourced from; e.g. in most models ET cannot access baseflow stores. It is noted that ET measurements are rarely used to calibrate rainfall-runoff models (with some exceptions, e.g. Frost et al., 2015).

Analysis of dynamics on multiple timescales may improve understanding of ET. Year-to-year variability in water balance may yield insights, for example into transpiration response to drought, as discussed below (Troch et al., 2009). Seasonal analysis may also provide understanding, since the amount of water removed via ET over a dry season may correlate strongly with the amount of water absorbed by a catchment in the ‘wetting up’ period of the early wet season (Grayson et al., 1997; Western et al., 1999; Potter and Chiew, 2009). Practically, the order of equations in a hydrologic model (i.e. is ET calculated before or after ‘effective runoff’?) also influences how much water is considered to be available for ET.

Most models do not differentiate between components of ET such as interception, transpiration and bare soil evaporation. The tendency of hydrologists to lump such processes together, both conceptually and in numerical models (e.g. Jakeman and Hornberger, 1993; Perrin et al., 2003) has been criticised in the literature (e.g. Savenije, 2004) and may impede meaningful diagnosis of model error. On a global scale, transpiration dominates ET, accounting for 60-90% (Jasechko et al., 2013; Schlesinger and Jasechko, 2014), so the next subsection considers vegetation dynamics specifically.
Vegetation dynamics

Could changes in vegetative water use have caused the poor DSST performance reported in Section 2.2.2? Despite many impacts of drier climatic conditions on vegetation, a compelling case is difficult to identify. For example, tree mortality (Allen et al., 2010; Park Williams et al., 2013) would be expected to decrease rather than increase PET, and climate-induced changes in species mix (eg. Hughes, 2003) would be unlikely to explain flow reductions on the timescales reported (eg. 13 years for the Millennium Drought). Reduced flows may occur during wildfire regeneration (Kuczera, 1987), but this is unlikely to be the sole cause because most studies used large samples of catchments that included many unburnt areas. Ukkola et al. (2015) linked reduced runoff in sub-humid and semi-arid catchments with recent changes to atmospheric CO₂, and Ajami et al. (2017) also suggested that observed non-stationarity of Australian hydrologic response might be linked to vegetation, and possibly to CO₂ fertilisation effects. However, the review by Peel (2009) reveals a lack of consensus on future CO₂ impacts on total vegetation water use. Troch et al. (2009) reported that vegetative water use is more efficient in drier conditions, meaning that vegetation uses a greater proportion of available water in dry years, consistent with the DSST results. These factors are not usually accounted for in conceptual rainfall-runoff models, which generally have very simplistic ET equations. Thus, some authors advocate the use of eco-hydrological models that include vegetation dynamics more explicitly, arguing that accounting for these dynamics leads to more defensible runoff projections (Naseem et al., 2015, 2016; Tang et al., 2018).

Water for transpiration is taken from the soil and substrate surrounding plant roots. The depth of this rooting zone varies in space and is thought to depend on geology, climate, and plant species mix (Milly, 1994; Rodriguez-Iturbe et al., 1999; Gao et al., 2014; de Boer-Euser et al., 2016). Some authors describe this as ‘co-evolution’ whereby catchment physical characteristics influence vegetation, and vice versa through soil formation (Troch et al., 2015; Savinije and Hrachowitz, 2016). The rooting depth provides water storage as a buffer against drought (Wang-Erlandsson
et al., 2016; Saft et al., 2016a), and understanding how this rooting zone evolves with climate (Troch et al., 2015) might improve hydrological projections. However, this evolution would only serve as an explanation for historic DSST results if it was shown to operate on similar timescales to the length of droughts used for DSST evaluation (eg. the 13 year duration of the Millennium Drought).

Characterising the root zone as a single thickness or depth of storage, as is often done in conceptual models, belies the complex sources of transpired water. For example, working in western USA, Dawson and Ehleringer (1991) found that adult trees directly adjacent to streams tended not to use stream water. Despite roots that were spread throughout the soil profile, water for transpiration tended to be sourced from relatively deeper soil layers (see also Evaristo et al., 2015; Bowling et al., 2017). With access to deep water, transpiration is unlikely to be a linear function of water storage in the soil, and this is at odds with the simplistic ET representations of many rainfall-runoff models. To consider these issues further, the next subsection considers the notion of catchment storage more generally.

**Catchment storage and memory**

The overall amount of water stored in a landscape (in both groundwater and soils) is a key state variable of the catchment system, and is strongly related to fluxes such as streamflow (Mcnamara et al., 2011). Storage goes up following rainfall, and lowers thereafter due to evaporation, transpiration, and drainage. Thus, it would seem that the key determinant of storage is recent hydroclimatic conditions, where ‘recent’ means days, weeks or possibly months. However, is it possible for storage to be sensitive to conditions on longer timescales? For example, if this year is wet, should the same catchment storage (and flow response) be expected regardless of whether the preceding few years were wet or dry?

Literature suggests that runoff and catchment storage can be sensitive to long term conditions. Working in the USA’s Sacramento River Basin, Risbey and Entekhabi (1996) examined runoff following isolated dry years and multiyear droughts, and noted that the flow response in the latter was relatively more subdued. To
describe the basin’s sensitivity to past years, they use the phrase ‘hydrological memory’. In southern Australia, groundwater records show considerable hydrological memory, recording the cumulative effect of sustained dry conditions. In the southwest, groundwater levels in some upland catchments have been in continuous decline since the 1970s following downward shifts in rainfall (Hughes et al., 2012, Section 2.1.6). Chiew et al. (2014) showed similar results in upland catchments in the southeast, with groundwater declines of 4 meters in an upland catchment. Vegetation appears able to access the declining groundwater store, since large-scale tree mortality has not been reported despite the dry conditions. In both cases, the declining groundwater levels were associated with significant streamflow decline.

**Deficit-limited models**

Despite their diversity, a common feature of most conceptual models is a ‘hard’ limit on water deficits (the amount of water that can be removed from storage). ‘Bucket’ models exemplify this: once the bucket empties, there is no water left, and ET ceases (eg. Beven and Freer, 2001b; Perrin et al., 2003; Chiew et al., 2002) although exceptions do exist (Jakeman and Hornberger, 1993; Evans and Jakeman, 1998). During long droughts, this conceptualisation may be unhelpful if simulated water storage becomes low enough to reduce ET to physically unrealistic values. The significant hydrological memory reported above may not be possible for a bucket model whose slowest-moving storage replicates baseflow recession. Since baseflow typically recedes in weeks or months, there is no mechanism for memory beyond a few months. Thus, the ‘bucket’ paradigm itself may hinder simulation of catchments with significant hydrological memory.

**Catchments as sets of spatially distributed interconnected storages**

Furthermore, conceptual models may be limited by their lumped nature. Most conceptual models are one dimensional, consisting of a set of buckets that are (notionally) vertically interconnected. Any spatial variability across the catchment is assumed to be implicit in the lumped model equations. In reality, spatial variability
is very important in hydrology (Grayson and Blöschl, 2001), and spatial theories such as the Variable Saturated Area concept have been very successful descriptions of rainfall-runoff response (Dunne, 1983; Quinn and Beven, 1993). In response, some conceptual models are formulated with spatial variability in mind but not explicitly represented (eg. Wood et al., 1992; Moore, 2007).

How could more explicit spatial representation aid model simulation of variable climatic conditions? Consider a catchment that has been subject to a long drought. Groundwater levels are low, soil moisture is depleted, and horizontal connectivity is minimal. Nonetheless, small amounts of streamflow continue (consistent with Hughes et al., 2012; Chiew et al., 2014), implying that groundwater levels are relatively close to the surface in some areas, such as valley bottoms. A wet year occurs (again, consistent with Chiew et al., 2014) and this is sufficient to saturate the valley bottoms. However, the wet conditions are insufficient to reduce deficits on all hillslopes to their pre-drought states, and portions of hillslopes remain hydrologically disconnected (Beven and Freer, 2001b), so that the runoff response is muted compared to pre-drought conditions (Risby and Entekhabi, 1996). Thus, different parts of the landscape retain different memory of past hydrometeorological conditions.

It may not be possible to simulate this situation using current conceptual models, particularly those with only one set of stores. If the wet conditions are sufficient to fill up all stores, memory of prior drought is completely removed. Lack of differentiation between different positions in the landscape (eg. valley bottoms versus hillslopes) may thus impede plausible characterisation of flow response.

The following quote by Petheram et al. (2011) relates to the above subsections:

*We propose that in the low-relief, moderate rainfall catchments of southern south-east Australia, relatively high groundwater levels may have amplified overland flow during pre-drought conditions by reducing the storage capacity of the unsaturated zone and by facilitating organised patterns of drainage and the connection of source areas of runoff as the soil wetted up during a rainfall event. Under a falling watertable, the storage capacity of the unsaturated zone increased, and hence saturation conditions were less likely to occur, and the connectivity of source areas was likely to*
2.4. **IMPROVING RAINFALL-RUNOFF MODEL STRUCTURES**

be less organised.

This explanation draws together the discussion of limited deficits (through mention of the falling water table) with the discussion of multiple storages (through the mention of connected, organised patterns of drainage). This confluence of factors seems to provide a plausible explanation for the poor DSST performance of conceptual models under drying conditions.

**Other possible causes of deficiencies**

Some hydrologists have suggested, in the case of the Millennium Drought, that the poor DSST performance may be due to a change in human interception activities, for example due to small private dams (*Petheram et al.*, 2011; *Potter et al.*, 2013). While there is evidence that such dams reduce streamflow in some catchments in this region (*Neal et al.*, 2001; *Schreider et al.*, 2002; *Fowler et al.*, 2015), the observed flow reductions are considered to be too large to be solely the result of farm dam interception due to the small overall volume of these storages (*cf.* *Nathan and Lowe*, 2012; *Sinclair Knight Merz et al.*, 2010).

Also, some authors have suggested that feedback could exist within the biophysical system that could lead to multiple stable states in catchment response. For example, *Peterson et al.* (2009) observed multiple stable states in a biophysical model arising from feedback between transpiration and groundwater head in a saline aquifer (*see also D’Odorico and Porporato*, 2004, for a hydrometerological example). If such stable states exist more generally in catchment systems, then a multi-year drought could cause the catchment to enter an alternate stable state, accounting for the observed persistent changes (*Saft et al.*, 2016a). However, multiple stable states arise through competing feedbacks, and a set of feedbacks that could explain the observed behaviour has not yet been identified.

**Summary**

In summary, conceptual rainfall-runoff models appear to underestimate ET during droughts. Conceptual models typically have simplistic ET formulations, and many
aspects of ET (particularly transpiration) may be characterised poorly. The continu-
uance of ET during long dry periods is impeded by the ‘bucket’ structure of most concep-
tual models. This hinders the accumulation of impacts of multiple-year dry periods, so that models cannot ‘remember’ preceding dry conditions for longer than a few months. Further, aspects of the runoff response related to spatial organisation
and hydrologic disconnection appear difficult to replicate with current conceptual models. Thus, there are numerous potential avenues of improvement for commonly used rainfall-runoff models.
Chapter 3

Research Questions

This chapter presents the research questions of this PhD and outlines how they were chosen. Section 3.1 compiles all research questions identified in the Literature Review. Section 3.2 outlines how the research questions were prioritised. Section 3.3 provides the final list of adopted research questions.

3.1 Potential research questions identified in Literature Review

The research questions arising from the literature review, in the order in which they arose, are:

i. Can ‘space-for-time’ methods provide superior hydrologic projections compared to common single-site approaches, in regions of high interannual variability such as southern Australia?

ii. Do monthly rainfall-runoff models provide more robust projections than daily models?

iii. When models fail the DSST, what aspects of the flow regime (signatures) are poorly matched?

iv. Is the poor performance of rainfall-runoff models over historic droughts due to poor calibration methods, model structural deficiencies, or both?
CHAPTER 3. RESEARCH QUESTIONS

Do any of the following calibration methods provide improved DSST performance, compared to least-squares-type approaches:

v Flow transformations that lessen the focus on high flow behaviour?

vi Matching targeted aspects of the flow regime (hydrologic signatures)?

vii Methods utilising the concept of data depth? and

viii Methods using metrics defined over a variety of timescales?

Do any of the following model structural changes provide improved DSST performance:

ix Less simplistic (possibly less linear) relationships between ET and water storage, to reflect vegetation’s access to water deep underground?

x Relaxing of the ‘hard’ limits on water deficits implied by bucket-type models, possibly being replaced by ‘softer’ systems of negative feedback?

xi More explicit representation of catchments as sets of spatially distributed interconnected storages, so that different parts of the landscape retain different memory of past hydrometeorological conditions?

3.2 Prioritisation of research questions

As outlined in Chapter 1, the research focus is improving rainfall-runoff simulations in changing climatic conditions, with two broad aims: to investigate causes of poor performance of rainfall-runoff models over changing climatic conditions; and to formulate changes to modelling practice that may improve performance of rainfall-runoff models over changing climatic conditions. These aims informed the selection of research questions.

Most of the above list falls into one of two categories: firstly, alternative calibration methods that may provide improved simulations in changing climates (items v-viii); and secondly, potential improvements to model structures (items ix-xi). An
3.2. PRIORITISATION OF RESEARCH QUESTIONS

An informed choice between these categories requires knowledge of whether, in general, the poor performance of rainfall-runoff models over historic droughts due to poor calibration methods, model structural deficiencies, or both (item iv). This was adopted as the first research question, worded as: Are current conceptual rainfall-runoff model structures deficient in their ability to simulate streamflow responses to long term changes in climate? The hypothesis for this research question was: the poor performance is due to poor or insufficient model calibration and evaluation techniques rather than deficient model structures. This research question was phrased as a ‘revisiting’ of the problem, since few researchers had previously explored the idea that the simulation deficiencies may be due to calibration methods rather than model structures.

As presented in Chapter 4, the answer to research question 1 was that both model calibration and deficient model structures are important in explaining the deficiency of simulations. In response, research question 2 focussed on improving calibration methods. To provide a manageable research question, some items were deemed out of scope. Firstly, the scope was limited to models commonly used in climate projections (daily conceptual rainfall-runoff models). This placed monthly models (item ii) out of scope. Secondly, the scope was limited to single-site calibration with commonly used inputs (precipitation and PET) and training data (runoff). This placed item i out of scope, because to provide hydrological projections at a given study catchment, this method requires data from many other catchments. However, items v-viii are within the scope of research question 2, which was phrased as: Which calibration metrics result in simulations that are robust to reductions in rainfall?

The final research question arose directly from research question 1. As mentioned, analysis addressing research question 1 concluded that the poor performance of rainfall-runoff models over historic droughts is due both to poor calibration methods and to model structural deficiencies. However, the cause varied depending on the catchment and model structure. Thus, in a case where a model has already failed split sample testing, a practitioner has no way of knowing a priori whether to focus on model structures or calibration methods. The existing literature provides no guid-
ance here (note, some frameworks in the literature (eg. Westra et al., 2014) do cover
data errors and model structures, but do not also cover calibration methods). Thus,
the final research question concerns this research gap. The question was phrased:
In cases where a model has already failed split sample testing, what general steps
should be followed to improve simulations under changing climates? As explained
in Chapter 7, this research question also partially addresses item iii, the diagnosis of
poorly matched aspects (hydrologic signatures) of the flow regime.

3.3 Final list of research questions

The final research questions are listed below:

Research Question 1

Question: Are current conceptual rainfall-runoff model structures deficient in their
ability to simulate streamflow responses to long term changes in climate?

Hypothesis: The poor performance is due to poor or insufficient model calibration
and evaluation techniques rather than deficient model structures.

Covered in: Chapter 4

Research Question 2

Question: Which calibration metrics result in simulations that are robust to reduc-
tions in rainfall?

Hypothesis: That there exists at least one metric with the property that optimising
that metric over a pre-change (wetter) period leads to improved model performance
over a subsequent post change (drier) period, compared to least-squares type meth-
ods.

Covered in: Chapters 5 and 6
Research Question 3

Question: In cases where a model has already failed split sample testing, what general steps should be followed to improve simulations under changing climates?

Covered in: Chapter 7
Chapter 4

Revisiting the apparent deficiency of conceptual rainfall-runoff models in changing climatic conditions

4.1 Chapter Summary

1 Hydrologic models have potential to be useful tools in planning for future climate variability. However, recent literature suggests that the current generation of conceptual rainfall-runoff models tend to underestimate the sensitivity of runoff to a given change in precipitation, leading to poor performance when evaluated over multi-year droughts. This research revisited this conclusion, investigating whether the observed poor performance could be due to insufficient model calibration and evaluation techniques. We applied an approach based on Pareto optimality to explore trade-offs between model performance in different climatic conditions. Five conceptual rainfall-runoff model structures were tested in 86 catchments in Australia, for a total of 430 Pareto analyses. The Pareto results were then compared

1This chapter has been published in Water Resources Research, as described in the preface
with results from a commonly used model calibration and evaluation method, the Differential Split Sample Test. We found that the latter often missed potentially promising parameter sets within a given model structure, giving a false negative impression of the capabilities of the model. This suggests that models may be more capable under changing climatic conditions than previously thought. Of the 282(347) cases of apparent model failure under the split sample test using the lower (higher) of two model performance criteria trialled, 155(120) were false negatives. We discuss potential causes of remaining model failures, including the role of data errors. Although the Pareto approach proved useful, our aim was not to suggest an alternative calibration strategy, but to critically assess existing methods of model calibration and evaluation. We recommend caution when interpreting split sample results.

The key points of this chapter are:

- Models may be more capable under changing climatic conditions than previously thought.
- Common calibration methods often fail to identify parameter sets that are robust.
- Caution is needed when interpreting the results of split sample testing.

4.2 Introduction

Water resource planning is essential to ensure the ongoing security of water supply for domestic, agricultural, industrial and environmental needs. Long-term streamflow projections inform this planning and help to anticipate potential future shortfalls in surface water supply. Estimates of water availability should take into account both historical observations of river flow and also potential changes in environmental conditions such as climate or land use.

Hydrologic processes exhibit variability and cyclical behaviour on a variety of timescales, from familiar short term cycles (diurnal, event and seasonal) to multi-decadal (Hurst, 1951). Alongside the reality of climate variability is the potential for
long term trends due to climate change (eg. Covey et al., 2003; Forster et al., 2007). A number of elements of the hydrologic cycle could be affected, including rainfall and evapotranspiration (Meehl et al., 2007; Donohue et al., 2010; Meville et al., 2012). Although the effects on precipitation are uncertain (Covey et al., 2003), many parts of the world, including southern Australia, are likely to see reduced rainfall (Chiew et al., 2009) and catchments may be persistently drier in the future than the past.

Hydrologic models are useful tools in planning for future variability in climate. They allow hydrologists to estimate the impact that long-term changes in climatic variables, such as rainfall, might have on water availability for human consumption or environmental needs. In this research we focus on conceptual rainfall-runoff models, which aim to represent mathematically the concepts underlying physical processes, without direct reference to physically based equations. Conceptual models generally have minimal data requirements, require minimal computing time, and often provide comparable simulations to more complex models (eg. Refsgaard and Knudsen, 1996), so they are relatively popular in practice. As reviewed below, many studies have concluded that conceptual models are generally not suitable when climatic conditions change (nevertheless they are often used in such conditions), and the intention of this chapter is to revisit this conclusion. Before reviewing this literature in detail we describe the tests that are commonly used to support the conclusion, specifically the concept of split sample testing.

To increase the level of confidence in the predictive capability of a given model, Klimes (1986) recommended a scheme known as the Split Sample Test, whereby a portion of historic recorded data are withheld from the calibration period, and used to check that the model can perform well over a period that it was not calibrated to—hereafter referred to as an evaluation period rather than using the common terms validation, verification or confirmation (Oreskes et al., 1994; Andrässian et al., 2009). In cases where a model will be applied in conditions different to the calibration period, Klimes (1986) suggested that the calibration and evaluation periods be specifically chosen so as to reflect a similar contrast in conditions, a test known as the Differential Split Sample Test (DSST). In the context of a changing climate, whereby rainfall
may be subject to long term trends, the DSST involves evaluating a model over a
period that is significantly drier or wetter than the calibration period. More recently,
variants of the DSST have been proposed, including the idea of using multiple cal-
ibration and evaluation periods via a sliding window in time (Coron et al., 2012,
2014; Thirel et al., 2015a).

Studies that have applied the DSST to assess the capabilities of models over
a changing climate have generally reported unfavourable results. Model predictive
ability following a change in climate does not appear to improve with more complex
models, as demonstrated by Refsgaard and Knudsen (1996) who tested three models
of varying complexity on three catchments in Zimbabwe. Furthermore, a number of
studies have identified significant bias following application of the DSST. Hartmann
and Bárdossy (2005) applied a lumped conceptual model to a 2000km² catchment in
Germany, calibrating successively to ‘wet’, ‘dry’, ‘warm’ and ‘cold’ years. They found
that models calibrated to the wet periods systematically overestimated flow during
dry periods unless the objective function explicitly included performance measures
calculated over longer (eg. annual) timesteps. Coron et al. (2012) applied three
conceptual models to 216 catchments and reported that “calibration over a wetter
(drier) climate than the validation climate leads to an overestimation (underesti-
mation) of the mean simulated runoff” (ibid. p1). Chiew et al. (2009) applied two
conceptual models to provide climate change projections based on downscaled GCM
outputs across south east Australia. Testing model performance over various peri-
ods with different climatic characteristics, they reported reductions in Nash Sutcliffe
Efficiency (NSE) value of 0.1 - 0.3 compared to the calibration period, and long term
bias of 30-40% in some cases. The recent workshop entitled Testing simulation and
forecasting models in non-stationary conditions (Thirel et al., 2015b), held under the
auspices of the International Association of Hydrological Sciences (IAHS), further
confirmed - for a wide range of models and catchments - that hydrological mod-
els tend to perform poorly if applied under changing climatic conditions (see Thirel
et al., 2015a, and citations therein).

Some researchers have sought to quantify acceptable changes in climatic vari-
4.2. **INTRODUCTION**

ables such as rainfall, such that a calibrated model still provides acceptable results. Vaze et al. (2010) tested four rainfall-runoff models in 61 catchments in South East Australia, and reported that the calibrated parameter sets generally gave acceptable simulations provided rainfall changes were not too large - no more than 15% less or 20% greater than rainfall over the calibration period. Similarly, (Singh et al., 2011) identified an acceptable change of 10% drier or 20% wetter for five catchments across the continental USA.

Other studies have phrased the problem in terms of the non-stationarity of model parameters across different climatic conditions. Merz et al. (2011) applied the HBV model to 273 catchments in Austria and found that parameters relating to snow melt and the nonlinearity of runoff generation tended to change with time, showing significant correlation with climatic variables such as temperature. Coron et al. (2014) similarly observed problems with parameter robustness in twenty mountainous catchments in southern France. Some studies have observed that even if a rainfall-runoff model may appear to perform poorly in the DSST, it is usually possible to find a parameter set that can match a given period, even if it is unusually dry or wet, provided that the model is directly calibrated to that period exclusively. This observation led to Li et al. (2012) recommending that “if a hydrological model is set up to simulate runoff for a wet climate scenario then it should be calibrated on a wet segment of the historic record, and similarly a dry segment should be used for a dry climate scenario” (ibid. p1239). Similar sentiments were expressed by Vaze et al. (2010). However, this solution is limited to providing projections that are within the range of climatic conditions experienced in the past (cf. Refsgaard et al., 2014). Choi and Beven (2007) tested a hydrologic model in a South Korean catchment and evaluated it over a variety of climatic conditions. Despite good performance according to classical performance measures on the timeseries as a whole, no parameter set tested was considered behavioural over all 15 of their categories of climatic conditions.

Despite these problems, some studies have had success searching for robust parameter sets, that is, parameter sets that can replicate streamflow over a wide variety of climatic conditions. Hartmann and Bárdossy (2005) formulated a number
of objective functions based on least squares calculations at different timesteps (e.g. daily, annual, decadal). Methods that combined both annual and daily objective functions into a single ‘meta-objective’ were shown to reduce the error in annual flows from 30% to 10%. Shamir et al. (2005a) applied similar multi-timescale logic but based their analysis on flow statistics (signatures) rather than least squares measures. The result was an ensemble of parameter sets that performed well on all timescales considered; the identifiability of parameters in the Sacramento model was also improved. Bárdossy and Singh (2008) introduced the statistical concept of data depth to hydrological modelling. A parameter set has greater depth if it is located closer to the centre of a cloud of well performing sets. They found that parameter sets with greater data depth were more robust in split sample tests and less sensitive to random errors in input data.

Although in the above discussion we have used the term ‘model’ quite loosely, henceforth we adopt the terminology outlined in Andréassian et al. (2009) where ‘model structure’ refers to a set of equations representing a catchment whereas the term ‘model’ refers to a model structure populated with a particular set of model parameters. A number of studies have concluded that a particular model structure is unsuitable for modelling under a changing climate (e.g. Vaze et al., 2010). Others have suggested that a given model structure needs changing to do so (Merz et al., 2011) or have gone further and actually produced a model structure specifically designed to simulate under changing climatic conditions (e.g. Ramchurn, 2012; Hughes et al., 2013). However, given the success of the studies mentioned above in finding more robust parameter sets under changing climates, perhaps the greater part of the problem lies with calibration and evaluation techniques rather than model structures. We suggest that a conclusion of model structure invalidity actually requires a much higher standard of proof than the tests of model evaluation suggested by Klemes (1986). To conclude that a model structure is invalid is to assert that no suitable parameter combinations exist (e.g. Vogel and Sankarasubramanian, 2003); whereas Klemes’ (1986) methodology seeks only to test the suitability of a chosen parameter set(s).
4.2. INTRODUCTION

This research sought to investigate the apparent deficiency of a range of conceptual rainfall-runoff model structures, across a large sample of catchments. The key research question was, Are current conceptual rainfall-runoff model structures deficient in their ability to simulate streamflow responses to long term changes in climate? As described above, some existing literature portrays rainfall-runoff models as suffering from poor performance if applied in climatic conditions different to those against which they were calibrated. The hypothesis tested here is that the poor performance is due to poor or insufficient model calibration and evaluation techniques rather than deficient model structures.

To conclude this section, we wish to clarify our intended meaning when using words such as deficient. Gupta et al. (2012) among others note that different hydrologists have different perspectives when defining model adequacy, contrasting the “physical science” viewpoint (where adequacy means consistency with the physical system) with the “engineering” viewpoint (where adequacy means that the model can emulate system input-output behaviour). In the context of rainfall-runoff models, a physical science viewpoint would insist that a model can realistically represent the dominant physical processes occurring in a river catchment, whereas an engineering viewpoint would focus on whether the model streamflow outputs match with observations. We affirm the physical science viewpoint and the need to advance hydrologic science by developing more physically realistic models. However, the general nature of our research question requires testing a large variety of case studies (86 catchments, 5 model structures, see Sections 4.3.3 - 4.3.4 and cf. Gupta et al., 2014), which renders detailed consideration of physical processes in each individual case difficult. Therefore, we use the word ‘deficiency’ in a sense consistent with the engineering viewpoint, and this study investigates the ability of models to provide an empirical match with observed streamflow data, in catchments subject to long-term changes in climate.
4.3 Method

4.3.1 Rationale

To explain the methodology, let us consider a simple hypothetical case study. A rainfall-runoff model structure A is applied to a catchment B using a calibration method C. Let us assume that method C is a single objective optimisation, such as would commonly be used within a DSST, optimising to a single objective function that varies between 0 (poor) and 1 (good). In order to conduct a DSST, the observed data are split so as to reserve a period for independent evaluation. The evaluation period is much drier, on average, than the calibration period. The result is a very good score over the calibration period (say, 0.9) but a very poor score over the drier evaluation period (say, 0.2). Since the purpose of the exercise is to identify a model that performs well in evaluation (Klemeš, 1986), it is tempting to conclude that model structure A has failed for catchment B, or that poor data quality is degrading performance. However, consider Figure 4.1. The parameter set identified in optimization lies in the red hashed region, whereas a solution in the blue dotted region is desired. However, the fact that the parameter set identified as optimal over the calibration period is in the red region, does not imply that no parameter set exists in the blue region. For example, it may be that the model structure itself is capable of simulating well in evaluation, but the relevant parameters remained poorly identified in this particular calibration exercise. Some other parameter set which has slightly lower (but still good) performance over the calibration period may exist that also performs adequately in evaluation. This latter question remains untested in this hypothetical case, and is the subject of this chapter. Note also that a parameter set in the red region may result from a calibration procedure caught in a local optimum (see eg. Arsenault et al., 2014).

Continuing the hypothetical case study, the periods are now switched, and the period that was the evaluation period now becomes the calibration period, and vice versa. The calibration is re-run and the result is that the score over the dry period is much increased (say, to 0.8) at the cost of some performance over the
nondry period (say, 0.5). In summary, in this hypothetical we have conducted two separate calibrations to two independent periods, and in each case we have obtained a parameter set that performs well over its training data, but poorly over the evaluation data. Figure 4.2a considers this as a two dimensional plot. Since both periods have had a turn at being calibration and evaluation periods, we dispense with this language altogether, and name them simply Period 1 and Period 2. For clarity we use the descriptions ‘nondry’ and ‘dry’, respectively. If parameter sets which are robust to changes in climatic conditions exist, they will have high values on both the x axis and the y axis. Whether or not our rainfall-runoff model structure A meets this condition depends upon the shape of the line that joins the two points and describes the trade-off between one objective and the other - the Pareto Front (Pareto et al., 1972). Under this scheme, a model structure with Pareto Front $\alpha$ in Figure 4.2b is more likely to produce robust simulations under changing climate than a model structure with Pareto Front $\beta$. Each Pareto Front is composed of numerous parameter sets, and Pareto Front $\alpha$ indicates robust simulations are possible because it contains parameter set(s) that have good performance on both the x axis and the y axis (eg. [0.8, 0.75]). Such robust parameter set(s) are akin (but not identical) to the
Figure 4.2: (a) Results of two hypothetical calibrations, plotted in two dimensional objective space. (b) Two Pareto Fronts joining the two points from (a). Each front is composed of multiple parameter sets. Obtaining curve $\alpha$ would demonstrate that a model structure has relatively greater potential for simulation under changing climatic conditions than would curve $\beta$.

hydrologic optimum of (Andréassian et al., 2012), which they define as the parameter set “that ideally would permit representing the catchment under all possible calibration periods encompassing climate forcings of interest, i.e. one allowing extrapolation”. In contrast, the two endpoints of the curve are the mathematically optimum parameter set(s) obtained via optimisation to each of the single objectives in turn.

Based on the above, the search for parameter sets that are robust to changing climatic conditions can be informed if we know the shape of a model structure’s Pareto Front. In this study we therefore applied a multi-objective optimiser to define the Pareto Front. Note that the method is intended only to critically assess existing methods of model calibration and evaluation; in this chapter we are not suggesting that this method should be adopted for general use in rainfall-runoff model calibration.

The remainder of this section is organised as follows: we first present the method for identifying the Pareto Front, called AMALGAM (Vrugt and Robinson, 2007); we then present the rainfall-runoff model structures to be tested; the catchment case studies; input data; the objective functions used; and methods for checking
the results of the AMALGAM algorithm.

4.3. METHOD

4.3.2 Multi-criteria analysis and Pareto search method

Multi-criteria analysis has been used in hydrology for some time in various contexts (Efstratiadis and Koutsoyiannis, 2010). Early examples included optimising treatment and monitoring of groundwater contamination (Cieniauski et al., 1995; Ritzel et al., 1994) and the incorporation of multi-response data in hydrologic modelling (Seibert, 2000; Madsen, 2003). Some authors adopted multi-objective approaches to improve identifiability of highly parameterised distributed models (e.g. Muleta and Nicklow, 2005; Bekele and Nicklow, 2007; Wöhling et al., 2013). Other studies have used multi-criteria approaches to integrate different data types into model calibration, including ‘soft’ information (such as local or expert knowledge, Seibert and McDonnell, 2002) and regionalised information (Kim and Lee, 2014). Gupta et al. (1998) suggested the potential for multi-objective calibration of rainfall-runoff models using different aspects of the same observed timeseries of flow (e.g. high flow versus low flow metrics), and a number of studies have adopted this approach (e.g. Booij and Krol, 2010; Kollat et al., 2012). The use of hydrologic signatures in model calibration can be seen as a variant on the multi-criteria approach, although methodological approaches vary (e.g. Shamir et al., 2005a; Yadav et al., 2007; Bárdossy, 2007; Winsemius et al., 2009; Vrugt and Sadegh, 2013). Gharari et al. (2013) noted that in addition to trade-offs between different metrics in the same time period, there are also trade-offs between model performance during one period and performance during another. They defined Pareto Fronts on both of these levels, and then designed a meta-Pareto analysis to choose parameter sets that provided the best overall compromise on the objective functions considered, over all periods considered. A key difference with the current study is that they were proposing a new model calibration approach, whereas in the current study we are using Pareto analysis to critically assess existing methods of model calibration and evaluation.

A number of algorithms to search for Pareto fronts are available in the hydrologic literature with notable early contributions being the development of the
hydrology-specific multi-objective calibration algorithms MOCOM-UA (Yapo et al., 1998) and MOSCEM (Vrugt et al., 2003). However, the concept is used in many fields and numerous algorithms from outside the field of hydrology are potentially applicable (e.g. Storn and Price, 1997; Deb et al., 2002). Algorithms are generally evolutionary rather than gradient-based, and this led Vrugt and Robinson (2007) to suggest a hybrid approach whereby the evolutionary process is conducted not only between different model parameter sets, but also between different search algorithms. The resulting Pareto search meta-algorithm, called AMALGAM (A MultiAlgorithm, Genetically Adaptive Multi-objective method), calls upon four commonly used methods for multi-objective searches (NSGA-II - Deb et al., 2002; Particle Swarm Optimization (PSO) - Haario et al., 2001; Adaptive Metropolis Search (AMS) - Eberhart et al., 2001; and Differential Evolution (DE) - Storn et al., 1997). These search algorithms are run simultaneously during an AMALGAM run, and the evolution of the population of parameter sets is directed by a combination of the search algorithms, with the influence of each in proportion to its performance at that point in the search. Vrugt and Robinson (2007) reported efficiency gains of up to a factor of 10 in some multi-objective problems. For this research, we adopted AMALGAM to search for Pareto Fronts.

4.3.3 Rainfall-runoff model structures

The intention of this study is to test a variety of model structures chosen to reflect common usage in the study area and, where possible, breadth of design of conceptual rainfall-runoff models. Since this study is focused in Australia three model structures that are commonly used in Australia were selected: GR4J (Perrin et al., 2003); SIMHYD (Chiew et al., 2002); and IHACRES (Jakeman and Hornberger, 1993; Ye et al., 1997). We adopt the version of IHACRES used in similar studies in Australia (e.g. Vaze et al., 2010) which incorporates the two parallel storages of IHACRES ‘Classic’ (Jakeman and Hornberger, 1993, see also Jakeman et al., 1990) with the option for a threshold of runoff production proposed by Ye et al. (1997). These three model structures, GR4J, SIMHYD and IHACRES are the result of three different
4.3. METHOD

ways of formulating conceptual rainfall-runoff models, as follows: (1) SIMHYD is an attempt to represent physical processes in conceptual equations, so that it has separate components for such processes as interception, infiltration excess overland flow, interflow/saturation excess flow and baseflow (Porter and McMahon, 1976; Chiew and McMahon, 1994; Chiew et al., 2002); (2) IHACRES has much less emphasis on physical processes, having been derived from mathematical analysis of the number of parameters that could reasonably be inferred from typical calibration data (Jakeman et al., 1990; Jakeman and Hornberger, 1993); and (3) GR4J has a similarly low emphasis on physical processes but was derived using an empirical approach that tested a large number of candidate structures and used a rejection method based on the empirical match with calibration data (Perrin et al., 2001, 2003). We consider that these three approaches to model formulation cover the majority of conceptual rainfall-runoff models currently in the literature.

In addition, two further model structures were included. GR4JMOD (Hughes et al., 2013) was chosen as a case study for improvement of rainfall-runoff models. Hughes et al. (2013) started with the GR4J model (Perrin et al., 2003) and tested a number of changes designed to better simulate environments with long-term (ie. multi-year) catchment storage. Their changes allowed the soil moisture to deplete below the level required for runoff production, effectively increasing catchment ‘memory’. They also added exponents to increase non-linearity of runoff production and actual evapotranspiration. Note that Hughes et al.’s 2013 module to account for changes in Leaf Area Index was not adopted here. Lastly, one model structure has been selected because it is widely used in the literature and in practice in the USA, namely SACRAMENTO (Burnash et al., 1973).

These model structures are summarised in Table 4.1. Model complexity varied, with the number of conceptual storages ranging from two to four, and the number of parameters ranging from four to sixteen. All models take the same inputs, namely, rainfall and potential evapotranspiration (PET - note the adopted version of IHACRES used PET rather than temperature). A lumped modelling approach was taken, whereby a single timeseries was derived for rainfall and PET in each
catchment (Section 4.3.5). The modelling framework was implemented in a hybrid Matlab-Fortran system whereby the rainfall-runoff models were run in Fortran 95 (checking against the code of the original authors where available - Table 1) which was called by the AMALGAM code in MATLAB provided by Vrugt and Robinson (2007).

4.3.4 Study area

This study was conducted in 86 catchments in southern and eastern Australia, as was shown in the Introduction Chapter, Figure 2.1. This region is well-suited to studying hydrological responses to long-term shifts in climate, because the variability of annual flows is relatively high on a global scale (Peel et al., 2001) and there have been a number of dry periods lasting several years or even decades on which to test model simulations. For example, the reduction in rainfall since the 1970s in the south west corner of Australia relative to the 1960s (eg. Petrone et al., 2010) has led local water authorities to run their long-term planning simulations using post-1975 data only. The south-east of the country experienced a severe and prolonged drought throughout much of the 2000s, known as the Millennium Drought (Potter et al.,
2010). River flows during the Millenium Drought, even given the low rainfall, were unexpectedly low in some areas (Potter and Chiew, 2011; Chiew et al., 2014; Saft et al., 2015). These droughts had numerous impacts on Australian society, including installation of alternative water sources such as desalination in most major cities, the cessation of irrigation in some areas causing changes in rural communities, and revision of water allocation arrangements to include water trading and provision for environmental flows (see eg. Aghakouchak et al., 2014).

The 86 study catchments were chosen from a wider set of ‘Hydrologic Reference Stations’ (Turner, 2012) defined by Australia’s Bureau of Meteorology as a set of catchments “with minimal water resource development and land use disturbances” (ibid, p6) such as regulation from large reservoirs and broadscale land use changes. Of the 154 Hydrologic Reference Stations that lie within southern Australia, (broadly defined as south of the Tropic of Capricorn), the list was refined according to:

i Data quality checking including inspection of quality flags, missing data, plotted daily data, inspection of double mass curves for flow and rainfall, and plotting long term climatic averages on axes similar to those used by Budyko (1974) - specifically, Actual Evapotranspiration versus Potential Evapotranspiration, both normalised by rainfall (see also Zhang et al., 2001).

ii Rain gauge coverage: Catchments were checked for coverage of rainfall gauges, and catchments with relatively low coverage were flagged.

iii Spatial rainfall contrasts: As mentioned above, a spatially lumped modelling approach was adopted, meaning that a single rainfall timeseries was used over a catchment (namely, the spatial average). Catchments with high spatial contrasts in rainfall are more difficult to simulate using a lumped approach, because the average rainfall is generally less representative of the rainfall extremes within the catchment. While a certain degree of rainfall contrast is usually inevitable due to topographic differences, the catchments with relatively higher contrast were flagged. Rainfall contrasts were assessed using the gridded rainfall data as described in the next section.
The final dataset of 86 catchments (Figure 2.1) was chosen so as to exclude those catchments with the clearest data issues, the lowest rain gauge coverage, and/or the highest spatial rainfall contrast, while aiming to preserve both the majority of catchments, and the spatial and climatic coverage inherent in the original dataset.

The set of 86 catchments vary in size from 4.4km$^2$ to 1106km$^2$, with 49 of the catchments between 100 and 500km$^2$ (see Figure 4.3). All of the catchments are in the temperate climate zone, falling within Group C of the Köppen-Geiger climate classification scheme (Peel et al., 2007). This means that the average maximum temperature of the hottest month is greater than 10$^\circ$ Celsius, and the average maximum temperature of the coldest month is between 0$^\circ$ and 18$^\circ$ Celsius. Mean annual rainfall is generally less than 1200 mm/year, while catchment average slope is generally less than 25% (Figure 4.3). Forest cover is generally high, with tree cover exceeding 90% in over half of the catchments. Catchment elevation ranges from sea level to 2000m AHD, although most catchments do not exceed 1500m AHD. Winter snowfall occurs in some catchments, but the snowpack is generally not sufficient to significantly affect hydrology. The development of small private waterbodies (referred to as ‘farm ponds’ in the USA and ‘farm dams’ in Australia) was also assessed where available. Over half of the catchments had an estimated farm dam storage of less than 5 ML/km$^2$, which can be considered quite low (Nathan and Lowe, 2012), although three catchments had more than 20 ML/km$^2$. These physical properties will be related to model performance later in the chapter.

4.3.5 Input data

The two main inputs to the rainfall-runoff models were rainfall and potential evapotranspiration (PET), each derived as a timeseries on a daily timestep. Rainfall was derived from the interpolated gridded product of Jones et al. (2009) which is available as a set of daily grids at a resolution of approximately 5km, based on gauged rainfall data and including land elevation as a spatial co-variate. For each day to be simulated, the spatial average across the catchment was derived from the daily grids

from Jones et al. (2009). PET estimates were derived using the Wet Environment method from Morton (1983). Given the relatively low spatial variability of potential evapotranspiration, this was extracted for the catchment centroid only, from the gridded datasets produced by Jeffrey et al. (2001) 3.

In the case of both rainfall and PET, the catchment boundary was required in order to extract information from the gridded datasets. Catchment boundaries were derived using flow analysis on Shuttle Radar Topography Mission (SRTM) data on a grid size of 1-second (approximately thirty metres). The post processed version by Gallant et al. (2011) was used for the flow analysis 4, which was done in ESRI’s ArcHydro toolbox using the D8 method to define flow pathways.

Streamflow data for the Hydrologic Reference Stations are publicly available 5. Quality codes were inspected and periods with quality issues were excluded from the analysis. Since quality code systems are different for each state of Australia, the details of this checking depended on location.

4.3.6 Defining dry periods and wet periods

As described in Section 4.3.1, the intention of the Pareto analysis is to search for parameter sets within a given model structure that provide a favourable trade-off between performance in dry climatic conditions and performance in wet climatic conditions. There were two separate tasks in order to develop this logic into a working system: firstly, to define ‘dry periods’ and ‘wet periods’ more precisely (this section), and secondly to choose a single objective function as an indicator of model performance over a given period (described in the next section).

To define dry periods, it would be possible to simply select the driest X% of years (or months), regardless of where those years may fall in the historic record. The results would be a set of years that are not concurrent. However, one of the key aspects of the recent droughts in Australia was not only their severity but also

---

their length and persistence; the persistent dry conditions have been shown to be associated with lower than expected streamflow response (Petrone et al., 2010; Potter and Chiew, 2011; Hughes et al., 2012, 2013; Potter et al., 2013; Saft et al., 2015). Therefore, we focused on sequences of dry years in this research, with the intention of examining multi-year droughts. While a number of studies have proposed methods of defining drought (see eg. Mishra and Singh 2010; 2011 and citations therein) there is no single accepted method for doing so. In this study we opt for a relatively simple definition, where we define the ‘dry period’ to be the driest consecutive set of years of a given length in the historic record. Given that the Millennium Drought is generally considered to have lasted from 1997 to 2009 (Chiew et al., 2014), we considered adopting a length of 13 years, or alternatively a round figure such as 10 years. However, in some places the drought was punctuated by an average or wet year mid-way through an otherwise dry spell (eg. the year 2000 in the state of Victoria).

It was felt that such a year could dominate the calculation of performance metrics relative to the drier years that are the topic of interest. Therefore, it was decided to use a shorter period, specifically seven years, instead. Thus, the dry period for this chapter (also called ‘Period 2’) is defined as the driest set of seven consecutive years in the historic record. This is defined according to streamflow, not rainfall.

While it is possible to define a ‘wet period’ in a similar way, (ie. by identifying the wettest series of concurrent years in the historic series), we have elected to adopt a method similar to that described in the hypothetical in Section 4.3.1. We defined “Period 1” as all years in the historic record, apart from Period 2 - that is, Period 1 is the complement of Period 2. This definition meant that Period 1 contains the majority of the historic data. Since the intention of this chapter was to provide a critique of the single-objective calibration approach (ie. single objective calibration to the nondry period and subsequent evaluation to the dry period), it was logical to provide as much calibration data as possible to this approach, such that the method under scrutiny was given the best possible chance to succeed. In any case, when calibrating to objective functions such as the Kling Gupta Efficiency used in this chapter (Gupta et al., 2009, see next section), the wetter periods tend to be matched
Figure 4.3: Catchment properties for the 86 study catchments. The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whisker are marked as outliers and are denoted as +. Catchment average slope was derived based on analysis of a DEM (Section 4.3.5) and represents the spatial average of cell-by-cell slope values. Forest cover was from Lymburner et al. (2011) and is the sum of the four landuses in the ‘tree’ category. Farm dam development is based on the dam locations and estimated volumes published by Department of Environment, Land, Water and Planning (2015a,b).

preferentially since the components of the KGE (linear correlation, error in mean and error in standard deviation) tend to be more strongly influenced by larger flow values. Thus, performance in Period 1 is an acceptable surrogate for performance over the wettest years in a given timeseries. For convenience, Period 1 will be referred to as the ‘nondry’ period.

We acknowledge that, in a given case study, Period 1 will usually contain some years that are relatively dry. Period 1 may contain entire sequences of droughts that were not the most severe on record, plus portions of the most severe drought not captured within Period 2 in cases where drought duration exceeds seven years. Conversely, Period 2 may contain years that were immediately prior to or following the drought of interest, in cases where the most severe drought is less than 7 years in duration. Nonetheless, these simple definitions were sufficient to examine differences in model performance between wet and dry periods, particularly for catchments where droughts tended to be longer and more severe.

Using these definitions, the flow series in each catchment was analysed so as to define Period 1 (nondry) and Period 2 (dry). As expected, the mean annual flow
Figure 4.4: (a) Boxplot of mean annual flows in Period 2 (dry period) expressed as a ratio to Period 1 (nondry period). The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whisker are marked as outliers and are denoted as + (b) histogram showing the most common starting years for the 7-year Period 2.

tended to be significantly less over Period 2 than over Period 1 (Figure 4.4a); for example, in over a quarter of catchments the flow reduction exceeded 70%. Figure 4.4b shows the distribution of start years over the set of 86 catchments. The duration of the Millennium Drought is considered to have been 1997 to 2009 (Chiew et al., 2014), so it is not surprising that the starting year of Period 2 is commonly within the range 2000-2004 (48 out of 86 catchments).

4.3.7 Objective functions

When deciding which objective function to use, the NSE (Nash and Sutcliffe, 1970) was the first candidate considered because its common use in practice means that values of NSE can be interpreted by a relatively wide audience. However, we encountered problems using the NSE. In some cases the NSE value was quite high (eg. 0.8) but upon further investigation the simulations were significantly biased. Gupta et al. (2009) provide an explanation for this in their decomposition of the NSE into the linear correlation and terms related to the error in the mean (ie. the bias) and error in the standard deviation. Gupta et al. (2009) noted that the bias term is normalised by the observed standard deviation, which means that in catchments with high flow
variability (as in this study) the magnitude of the bias can be high without penalising the NSE score. One option was to add a bias weighting to the NSE, as applied by, for example, Vaze et al. (2010). However, Gupta et al. (2009) noted a further problem with the treatment of the standard deviation $\sigma$ in the NSE, regarding the ratio $\sigma_{\text{simulated}}/\sigma_{\text{observed}}$. Although this ratio should ideally have a value of unity, the mathematically optimum value for NSE occurs when the ratio is equal to the linear correlation. Given these problems, this study adopted the alternative objective function proposed by Gupta et al. (2009), called the Kling-Gupta Efficiency, or KGE. The KGE is a function of the same three components as the NSE (linear correlation; error in mean; error standard deviation) but the formulation removes the interactions between the components, providing a more robust measure of model performance. For those readers who are not familiar with KGE scores, in the Appendix we provide a table that relates the KGE to the more familiar NSE objective function (Figure B.4), and also highlights the problems noted by Gupta et al. (2009).

4.3.8 Results checking

Since AMALGAM is an evolutionary algorithm, it is possible that calibration runs may proceed in different directions through the parameter space and have divergent end results (see, eg. Arsenault et al., 2014; Peterson and Western, 2014). To check the consistency of the AMALGAM results, we started with a relatively low number of function evaluations (10,000) and ran the algorithm three times, resulting in three different Pareto Fronts. These Pareto Fronts were checked for consistency both visually and using the numerical rule that the Euclidian Distance separating any two of the three curves could not exceed 0.01 at any point on the curves. If this numerical rule was violated, the number of function iterations was doubled and the analysis re-run and re-checked. Around one quarter of the case studies passed at the first iteration (ie. 10,000 function evaluations). Case studies that failed the numerical test at 40,000 iterations were manually (visually) checked and accepted only if the differences were judged to be immaterial to the conclusions of this chapter.
The presentation of results in the following section initially focuses on one objective (i.e., one period) at a time, before moving to consideration of the two objectives (i.e., performance over dry and nondry periods) simultaneously. Presentation of the AMALGAM results in this way implicitly assumes that AMALGAM is a sufficiently powerful search algorithm to find the optimum of a single objective. Another way of stating this assumption is that the endpoints of the Pareto Curves are assumed to be accurate. To test this assumption, the single-objective optimization algorithm CMA-ES (Hansen et al., 2003) was applied. CMA-ES has been widely used across a number of fields and tested favourably in the context of hydrology compared to more common methods in hydrology such as Shuffled Complex Evolution (Duan et al., 1992; see Arsenault et al., 2014 and Peterson and Western, 2014. In the current study, CMA-ES was trialled in ten catchments, for each of the five model structures, for each of the two objectives (KGE over Period 1 and KGE over Period 2). This gave a total of one hundred CMA-ES case studies. Similarly to AMALGAM, CMA-ES was run three separate times and if the results were not consistent, the number of restarts (the only user-defined parameter in CMA-ES) was increased by one (starting from zero restarts) and the process was repeated.

For brevity, the CMA-ES results are not shown in the body of this chapter but are provided in the Appendix (Figure B.5). In summary, the results indicated that AMALGAM was a capable and reliable optimizer to a single objective. Optimisation results (in terms of KGE scores) were within 0.005 in 76 of 100 cases. In the remaining 24 cases AMALGAM produced the best result in 15 and CMA-ES in 9. There were a few cases where AMALGAM results were significantly better than CMA-ES. In fact, ordering the case studies according to the absolute difference between the two results revealed that the top five cases (cases of greatest difference) were all cases where AMALGAM found a better solution than CMA-ES. We also note that, on average, the AMALGAM algorithm generally used less function evaluations than CMA-ES, although this varied based on the case study. Given these favourable results, we will now present the AMALGAM results with similar confidence as we would have in a dedicated single-objective optimizer.
4.4 Results

4.4.1 Performance when optimising to each objective in isolation

As demonstrated in the previous section, although a tool for multi-objective problems, the AMALGAM algorithm can also be used to provide results of a single-objective optimization, by considering the endpoints of the Pareto curves only. In this section, we present single-objective Differential Split Sample Test results, extracted from the wider set of AMALGAM outputs.

In general, optimising the rainfall-runoff models to KGE over Period 1 (nondry) provided good KGE values over Period 1 (Figure 4.5a). The median KGE score across all 86 catchments was 0.8 or higher, regardless of which rainfall-runoff model structure was chosen. For those readers who are not familiar with KGE scores, in the Appendix we provide a table that relates the KGE to the more familiar NSE objective function (Figure B.4). The GR4J and GR4JMOD model structures appeared to perform best. However, when the same parameter sets were evaluated by simulating flows over the driest 7 consecutive years (Period 2), model performance was much lower (Figure 4.5c). The model structures with the highest calibration KGE scores (GR4J and GR4JMOD) showed negative evaluation KGE values in more than 25% of catchments. IHACRES was comparatively better, with a median score of 0.67. Nonetheless, in general, the performance was markedly reduced when moving from wetter to drier climatic periods. Furthermore, some of the lowest values of KGE in evaluation corresponded to relatively high KGE values in calibration (Figure 4.6). These findings are consistent with the literature review (eg. Vaze et al., 2010; Coron et al., 2012; Thirel et al., 2015a).

If the dry period (Period 2) was used as the calibration period instead of the evaluation period, results demonstrated that the rainfall-runoff models are generally able to replicate the flows during dry conditions, provided they are directly calibrated to them in isolation. However, there were some exceptions, particularly for the GR4J model structure, as shown by the outliers in Figure 4.5d. The reduction in performance between the calibration period (dry period, Figure 4.5d) and the
CHAPTER 4. REVISITING MODEL DEFICIENCY IN CHANGING CLIMATE

evaluation period (nondry, Figure 4.5b) is less pronounced than in the previous case (Figure 4.5a/c) but is still evident. As above, some of the lowest values of KGE in evaluation corresponded to relatively high KGE values in calibration (Figure 4.6).

In summary, the model structures tested were generally able to replicate flows over a given set of climatic conditions, whether dry or wet, provided that they were directly calibrated to those conditions (Li et al., 2012). The key problem was that the parameter sets identified by optimization to one set of climatic conditions performed poorly in different conditions; that is, the mathematically optimum parameter sets identified were not robust to changes in climate. In subsequent discussion, the results presented in this section will be referred to as the results of a ‘single-objective DSST’, since the models were calibrated to only one objective at a time; i.e. KGE in one set of climatic conditions, with subsequent evaluation in different climatic conditions. These single-objective DSST results are used in this chapter as a baseline method representing common practice.

4.4.2 Pareto Curve results

For each of the five model structures, AMALGAM was applied to derive a Pareto Front between the two objectives (i.e. between KGE in Period 1 (nondry) and Period 2 (dry)) in each of the 86 study catchments. To explain and interpret these results, we first use the example of the Rocky River upstream of Gorge Falls (Station A5130501), a 190km² catchment on Kangaroo Island, South Australia (mean annual rainfall = 730 mm/year; rainfall-runoff ratio 0.1). The dry period in this catchment was found to be 2001-2007 inclusive, and average streamflow over this period was only 40% of the long-term average. For this station and the IHACRES model structure, the single-objective DSST metrics were:

- $\text{KGE}_{\text{nondry}}$ (calibration) = 0.835, $\text{KGE}_{\text{dry}}$ (evaluation) = 0.581;
- $\text{KGE}_{\text{dry}}$ (calibration) = 0.833, $\text{KGE}_{\text{nondry}}$ (evaluation) = 0.621.

These figures indicate that the single objective approach identified parameter sets that performed well in one set of climatic conditions or the other, but not both.
4.4. RESULTS

Figure 4.5: Values of Kling Gutpa Efficiency (KGE) for calibration and evaluation when optimized to Period 1, the nondry period (top) and Period 2, the dry period (bottom). Note that negative values exist but are not shown. The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whiskers are marked as outliers and are denoted as +.

Let us now consider whether the Pareto approach can identify robust parameter sets that perform well in both periods. Figure 4.7 shows the Pareto Front identified by AMALGAM, displayed in two-dimensional objective function performance space. The results quoted above form the endpoints of the Pareto curve in this space (ie. the endpoints are [0.581, 0.835] and [0.833, 0.621]). Since AMALGAM is an evolutionary method that uses a finite population, the front is displayed not as a continuous line but as a set of discrete points, one for each parameter set (in this case N = 100). A number of those parameter sets are in the region of the objective space
where $KGE_{\text{dry}}$ and $KGE_{\text{nondry}}$ both exceed 0.8. Thus, given that the values of both objectives are favourable for the same parameter set, we may cautiously conclude that the IHACRES model structure is capable of providing robust simulations over changing climatic conditions (assuming that the KGE can be considered a suitable indicator of simulation performance). Henceforth in this chapter we will use the terminology ‘false negative’ to refer to cases such as this where suitable parameter set(s) exist within a model structure, but the DSST fails to find them.

Next, let us consider the results for other model structures applied to the same catchment. While one model structure (SACRAMENTO) performed better, the remainder did not, as shown in Figure 4.8. The GR4J and GR4JMOD structures were capable of high KGE scores in either the dry period or the nondry period, as indicated by the end-points of the Pareto curves (cf. Figure 4.5). However, the curves joining these points do not approach the region of favourable trade-off mentioned above; that is, there were no parameter sets robust to changes in climate for these structures in this catchment. Note that in Figure 4.8, the individual markers have been replaced by lines for ease of viewing.

The results for GR4J and GR4JMOD in Figure 4.8 provide an instructive case study in model assessment. The endpoints of the curves are similarly placed for each of these two models. Thus, use of a single-objective DSST (as presented in the previous section) would lead to the erroneous conclusion that the alterations to GR4JMOD by Hughes et al. (2013) made negligible difference to the model’s capabilities. In contrast, Figure 4.8 shows that this is not the case by the divergence of the purple GR4JMOD curve from the orange GR4J curve. Although the difference in this case is relatively modest, it is not an isolated case - four further case studies are shown in the Appendix (Figure B.6). Thus, use of the single-objective Differential Split Sample Test may result in situations where highly successful model improvements are discarded as ineffective.
Figure 4.6: Scatter plots of calibration versus evaluation KGE values when calibrating to Period 1, the nondry period (top) and Period 2, the dry period (bottom). Each circle represents a catchment. Values of calibration KGE scores (x axis) versus evaluation KGE scores (y axis) for the same parameter set. Note that negative values exist but are not shown.
Figure 4.7: Pareto Front identified by AMALGAM between the two objectives, for the Rocky River upstream of Gorge Falls (A5130501), using the IHACRES rainfall-runoff model structure.

Figure 4.8: Pareto curves for each model structure for the Rocky River upstream of Gorge Falls (A5130501).
4.4. RESULTS

4.4.3 Identifying model structures that meet modelling standards

One difficulty in moving from a single catchment example to the full set of 86 catchments is the challenge of displaying the results meaningfully across such a large sample. For the interested reader, the Pareto curves for every combination of model structure and catchment are provided in full in the Appendix, Figures B.2 and B.3. Although we experimented (not shown) with measures to characterise the shape of the Pareto curve, we here focus instead on whether or not a given rainfall-runoff model structure is capable of robust simulations under changing climatic conditions, as indicated by high KGE values. Graphically, such model structures have Pareto Curves that contain parameter sets that are relatively close to the ‘perfect’ point, [1, 1].

Although it is difficult to say exactly how close is ‘sufficient’ for a given case study, for the present study it is useful to define some subjective performance standards. By defining what ‘success’ is (albeit in a subjective fashion), these standards allow us to more easily summarise the skill of the Differential Split Sample Test in identifying ‘successful’ model structures. Two attempts at defining such a standard are depicted in Figure 4.9. In Standard 1, a ‘successful’ model is one in which the model efficiency (KGE) at some point on the Pareto Curve exceeds 0.7 in both the dry and nondry periods. Standard 2 is similar except that the KGE benchmark is now 0.8; a higher standard of performance. Many other different standards could be formulated, and it is expected that the most suitable standard may depend upon the particular objectives of the study at hand. For the purposes of interpreting results in this chapter, we will proceed with these two standards, and denote any parameter set that meets a given standard to be ‘suitable’ (note that concepts of model adequacy are discussed in Section 4.5.2).

For each case study (ie. combination of model structure and catchment) we now ask two questions:

i. Would a suitable parameter set be found by a single-objective DSST calibrating over the nondry period (y-axis) and evaluating over the dry period (x axis)?
(Note that this corresponds to the left hand extreme of the Pareto Curve, such that the y-axis ordinate is maximised).

ii Would a suitable parameter set be found by AMALGAM? ie. is any portion of the Pareto Curve within the boxes of Figure 4.9?

We note that the DSST in point (1) above could equally be defined the other way around, with calibration over the dry period and evaluation over the nondry. However, climate projections for southern Australia generally agree that long-term average rainfall is likely to reduce under climate change (eg. Chiew et al., 2009). Thus, it is more relevant within this study area to evaluate models in conditions that are drier than the calibration period.

There are three possible combinations of answers to the above questions:

a Suitable parameter set(s), ie. parameter set(s) that meet the performance standard, were found by both the single-objective DSST and AMALGAM;

b Suitable parameter set(s) were not found by the single-objective DSST but were found by AMALGAM; and

c Suitable parameter set(s) were not found by either method.

To explain these categories graphically, consider the curves in Figure 4.9, which show Pareto results for Home Creek at Yaark (Station 405274, 181.6 km², mean annual rainfall = 744 mm/year; rainfall-runoff ratio 0.18). As an example, we consider the results for Standard 2 (dark grey). Only two of the model structures have a portion of the Pareto Curve within the box for Standard 2 - Sacramento (red) and IHACRES (green). This means that GR4J, GR4JMOD and SIMHYD are all in category (c) with respect to Standard 2. With respect to SACRAMENTO, although it can fulfil Standard 2, the parameter set that would be chosen by the single-objective DSST to Objective 1 (ie. the endpoint [0.51, 0.90]) does not fulfil Standard 2. Thus, SACRAMENTO is in category (b) with respect to Standard 2. For IHACRES, the parameter set that would be chosen by the single-objective DSST to Objective 1 (ie.
Figure 4.9: Pareto fronts for Catchment 405274, with annotations regarding the meeting of modelling standards 1 (light grey) and 2 (dark grey). The ticks and crosses refer to results for the single-objective DSST to Objective 1 (left of the line) and AMALGAM (right of the line).

the endpoint \([0.92, 0.88]\) does fulfil Standard 2. Thus, IHACRES is in category (a) with respect to Standard 2.

Hypothetically, if the results across all catchments and model structures indicated a dominance of case (a), then we would conclude that there are in fact few problems with current rainfall-runoff model structures simulating changing climates (in the ‘engineering’ sense; Section 4.5.2), although there might still be some scope to improve them. However, the results presented above (eg. Figure 4.5) have already demonstrated that this is not the case. Thus we are left with (b) or (c). Dominance of (b) would indicate that common single-objective calibration methods (as commonly used in the Differential Split Sample Test) generate an abundance of false negatives, and thus the problem is with the calibration methods, not with the model structures themselves. Dominance of (c) would support the argument that the model structures themselves need to be improved in order to provide an empirical match with streamflow data.

The results (Figure 4.10) depend on the modelling standard used, and on the model structure tested. Looking first at the lower of the two standards (Standard
1), for some model structures (eg. GR4J, GR4JMOD, SACRAMENTO) the cases are relatively evenly split between case (b) and case (c). This means that it was just as common for failure in a DSST to be the result of the calibration method as it was the result of the model structure. Thus, with regards to the hypothesis, both the models and the calibration methods need improvement in order to successfully model changing climatic conditions.

The IHACRES model structure once again provides an interesting case study. IHACRES was able to attain Standard 1 in a very high proportion of catchments: 74 out of 86 (i.e. 37 catchments in category (a) plus 37 catchments in category (b)). This would suggest that the model structure itself is relatively well suited to simulating changes in climate and does not require change to provide an empirical match with data. However, of the 74 that were successfully modelled by IHACRES, the single-objective DSST was only able to find a suitable parameter set in 37 cases (category a). The remainder (category b) were catchments where AMALGAM found a suitable parameter set but the single-objective DSST did not. This is not a particularly favourable success rate for the DSST, and suggests the need to review the use of single objective optimization methods in model calibration.

However, the interpretation shifts if Standard 2 is adopted instead of Standard 1. In this case the number of catchments where the modelling standard is not met is around 50% in the case of GR4JMOD and SACRAMENTO, and greater for GR4J and SIMHYD. The IHACRES model is able to meet this modelling standard in 59% of cases (24+27=51 out of 86 catchments) compared to 87% (74) for Standard 1. Thus, if this higher standard is adopted, one possible conclusion is that the current generation of rainfall-runoff model structures, including IHACRES, require improvements to simulate changes in climatic conditions in order to produce an empirical match with data. An alternative explanation is that the failure to attain the modelling standard is due to data errors (Section 4.5.3).

The pie charts to the far right of Figure 4.10 present the results in the case where a modeller is able to apply all five of the model structures to every catchment and has the freedom to adopt the best model whatever it may be. In this case,
suitable parameter sets are found during the single-objective Differential Split Sample Test in 53% of catchments (46 out of 86) for Standard 1, and 33% of catchments (28 out of 86) for Standard 2. There still remains a significant portion of catchments that are not modelled satisfactorily by any of the 5 model structures: 12 catchments out of 86 (14%) in the case of Standard 1, and 30 catchments out of 86 (35%) in the case of Standard 2.

4.4.4 Examination of catchments where models failed

We examined those catchments where the model structures failed to meet Standard 1 and/or Standard 2. Two main avenues were explored: firstly, we analysed the Pareto Curves and considered what the form of these curves may indicate about the type of model failure; and secondly we examined the physical and climatic properties of these catchments. For brevity, some elements of this discussion are summaries only,
CHAPTER 4. REVISITING MODEL DEFICIENCY IN CHANGING CLIMATE

with a reference to the appendices for more detail.

Having failed to meet the standard, every instance considered was one where no single parameter set could simulate flows satisfactorily in both wet and dry periods. However, we categorised failure instances further according to whether or not the model structure was able to meet the standard in a given objective when optimized to it in isolation. The results varied by model type: for example, GR4J and GR4JMOD were exceptionally good at meeting the modelling standards in a given objective provided that they were calibrated to it in isolation; i.e. the maximum possible value in each objective was high, but there was also a high degree of trade-off in between these endpoints. We categorised this type of failure with the phrase “Model structure can simulate both dry and wet periods, but not with the same parameter set”. In contrast, this type of failure was not common with IHACRES and SIMHYD, particularly for Standard 2, in which the category “Appears to be deficient in this catchment, regardless of climate” claimed the highest proportion of failures. The full results of this analysis are shown in the Appendix, Figure B.2.

Next, we focussed on the physical properties, including location, of the catchments where the model structures failed. For this analysis, we examined only those catchments where none of the model structures were able to meet the required standard. As per Figure 4.10, there were 12 such catchments in the case of Standard 1 (labelled “FF” since they failed both standards) and a further 18 in the case of Standard 2 (labelled “PF” since they passed one standard and failed the other).

In terms of geographic location, the instances of model failure are relatively well dispersed. In terms of failure to meet Standard 1 (red), there appeared to be two regions where model structures were more likely to fail: the central part of the state of Victoria, and the northern-most catchments tested in the state of Queensland. There were also a number of catchments failing Standard 2 (yellow) in the eastern highlands of New South Wales. A map is provided in the Appendix (Figure B.7).

Figure 4.11 shows the physical characteristics of catchments where model structures failed one or both standards. We selected five characteristics for testing, based on their perceived importance to hydrology: catchment area; rainfall; slope; forest
4.4. RESULTS

cover; and degree of development of private farm dams. Soil type and geology are also perceived to be important, but there are few high-quality national soil type / geology datasets that are numerical (ie. non-categorical). In addition to the five characteristics above, the observed severity of drought was also included, measured as the ratio of mean annual flows during the dry period to mean annual flows during the nondry period. From Figure 4.11a, catchment area appears to have little bearing on the failure of the model structures. However, Figure 4.11b and c show that cases of model structure failure tended to be in drier catchments, and where flow reductions during Period 2 were greatest.

To further investigate these results, we applied the non-parametric one-sided Rank-Sum Test, otherwise known as the Wilcoxon-Mann-Whitney test or the Mann-Whitney U test (as described by eg. Wilks, 2011; see also Wilcoxon, 1945 and Mann and Whitney, 1947). This evaluates the probability of the null hypothesis that two groups of data (in this case, characteristics of catchments where a modelling standard was / was not met, respectively) came from the same underlying distribution. By concentrating only on relative ranks rather than actual values, this test resists being influenced by one or two extreme values, which is important because some catchment characteristics have quite skewed distributions. The results (Table 4.2) confirmed that the catchments where the modelling standards were not met tended to be those with lower rainfall, lower slope and a greater relative reduction in flow during the seven driest consecutive years. Catchment area was less strongly related to modelling performance than these three, and forest cover less so again.

Since the group that failed Modelling Standard 1 (Group FF) is such a small sample size, we provide a catchment-by-catchment list of characteristics for each member of group FF in the Appendix Table B.1. Inspection of the individual characteristics of group FF reveals that although there appears to be differences between the boxplots for catchment average slope and forest cover, the reality is more complex, with group FF being spread across a relatively wide range in both cases.

In terms of farm dams, estimates of farm dam volume were only available for catchments in the State of Victoria (N = 39). Two of the three catchments where
Figure 4.11: Physical characteristics of catchments where model structures failed one or both standards. Four boxplots are shown for each characteristic: all catchments (marked A; N = 86); cases where both standards were passed (marked PP; N = 56); cases where Standard 1 was passed but Standard 2 was failed (marked PF; N = 18); and cases where neither standard was met (marked FF; N = 12). Farm dam data were only available for Victoria, so that the N values are different in plot f (NA = 39, NPP = 27, NPF = 6 and NFF = 6). The whiskers extend a maximum of 1.5 times the interquantile range. Values beyond the whiskers are marked as outliers and are denoted as +.

Table 4.2: Results from the non-parametric rank-sum test to test whether catchment characteristics differed between catchments where a given modelling standard was not met (by any model structure) and those where it was. Columns two and four indicate the probability that the observed differences in characteristics between the two groups of catchments arose purely by chance.

<table>
<thead>
<tr>
<th></th>
<th>Relating to standard 1</th>
<th></th>
<th>Relating to standard 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p value</td>
<td>Significant at 95% level?</td>
<td>p value</td>
<td>Significant at 95% level?</td>
</tr>
<tr>
<td>Catchment area</td>
<td>0.0953</td>
<td>no</td>
<td>0.0304</td>
<td>yes</td>
</tr>
<tr>
<td>Mean annual rainfall</td>
<td>0.0002</td>
<td>yes</td>
<td>&lt;</td>
<td>yes</td>
</tr>
<tr>
<td>Dry period flow ratio</td>
<td>0.0003</td>
<td>yes</td>
<td>&lt;</td>
<td>yes</td>
</tr>
<tr>
<td>Catchment average slope</td>
<td>0.0165</td>
<td>yes</td>
<td>0.0001</td>
<td>yes</td>
</tr>
<tr>
<td>Forest Cover</td>
<td>0.4161</td>
<td>no</td>
<td>0.2343</td>
<td>no</td>
</tr>
<tr>
<td>Farm Dam Development</td>
<td>0.0553</td>
<td>no</td>
<td>0.0041</td>
<td>yes</td>
</tr>
</tbody>
</table>
farm dam density exceeds 20 ML/km² were catchments where modelling standard
1 was not met, and it is possible that harvesting of water by farm dams in these
catchments is causing difficulties in modelling. However, the other catchments had
much lower levels of development of farm dams so it is unlikely that farm dams are
degrad ing model performance in these catchments. Further research is required
to investigate whether rainfall-runoff modelling in the two catchments with farm
dam density exceeding 20 ML/km² might be aided by quantification of farm dam
interception.

The results of this study are partially consistent with recent findings of Saft
et al. (2015) who analysed changes to the relationship between rainfall and runoff
on an annual timestep, in the same study area. They found that changes to the
relationship were more likely in drier catchments (upheld here) with low slope (upheld
here) and low forest cover (not upheld here, although the catchments used in this
study generally had greater forest cover than those in Saft et al. (2015). Note that
although bushfires are relatively common throughout Australia, we could not find any
evidence linking bushfire history with the failure of models to attain the modelling
standards (Section B.3).

4.5 Discussion

4.5.1 Results summary

Although the results above are specific to the catchments, data, models and objective
functions used, they are potentially relevant to any study that has rejected a model
structure based on a poor match with streamflows in an independent evaluation pe-
period (eg. a DSST). The results show that a significant proportion of such rejections
may be spurious because parameter sets may exist that fulfil a given set of perfor-
ance criteria but remain undetected during calibration. Thus, poor performance in
evaluation in a split sample test is a poor basis on which to reject a model hypothesis,
although it is adequate for rejecting the model/calibration method combination.

As noted in the method section, the Pareto framework used here was intended
only to critically assess existing methods of model calibration and evaluation; in this chapter we are not suggesting that the method should be adopted for use in rainfall-runoff model calibration. The reasons for this are explained below (Sections 4.5.2 and 4.5.3).

4.5.2 Getting the right answers for the wrong reasons?

We now consider whether or not a parameter set or model structure that is found to fulfil the adopted KGE performance criteria (ie. get the ‘right answer’) can be considered ‘adequate’ or ‘valid’. Firstly, it is widely acknowledged that one performance criteria (eg. KGE) is insufficient to ensure a holistic match with observed flows (Oudin et al., 2006; Gupta et al., 2009; Berthet et al., 2010; Andréassian et al., 2012), even if jointly considered over two contrasting periods as demonstrated here. As an example, consider the progression of modelling bias with time for the parameter sets shown in Figure 4.12. Even though long-term bias is a component of the KGE, the 10-year rolling average bias still deviates considerably from zero for the three parameter sets shown in each case, and shows some similarity with the results of Coron et al. (2014), particularly in catchment 613002. Choosing a parameter set that performs well in both periods (red) does not guarantee unbiased simulations over the modelling period, although GR4JMOD performs better in this aspect in the second case (A5040517) than the first (613002). This analysis of bias, based on the format of Coron et al. (2014), is shown for other selected case studies in the Appendix, Figures B.8 to B.11. From these examples it is clear that a high KGE score may mask underlying discrepancies in matching the observed data. Furthermore, even a near-perfect match with observed streamflows would not necessarily imply that a rainfall-runoff model is ‘adequate’ or ‘valid’, depending on the philosophical viewpoint. As discussed at the start of this Chapter, a near-perfect match with observed streamflows corresponds to adequacy in an operational or ‘engineering’ sense (Gupta et al., 2012) but a ‘physical science’ approach would ask whether the model is getting the right answers for the right reasons (Kirchner, 2006; Gupta et al., 2012). Under this viewpoint, models are adequate only if consistent with dominant physical
4.5. DISCUSSION

Figure 4.12: Long-term simulation bias (right) for three selected parameter sets (left), after Coron et al. (2014, Figure 5), for two selected case studies. Simulation bias is plotted as a ten-year moving average, and the ten-year moving average streamflows are also plotted for reference, in blue. The case study catchments are 613002 (Harvey River at Dingo Road, Western Australia; 147.5 km²; mean annual rainfall = 992 mm/year; rainfall-runoff ratio 0.21) and A5040517 (First Creek at Waterfall Gully, South Australia; 5.1 km²; mean annual rainfall = 992 mm/year; rainfall-runoff ratio 0.19). Despite similar positions in objective space for the red parameter set, changes in long term average streamflow are more faithfully tracked in A5040517 than 613002.

processes. As noted in the introduction, this is difficult to test in practice for a large sample of catchments, and thus we do not assess the adequacy of models in this physical science sense. Given that some processes that are thought to be important are not represented by the conceptual models used in this study (eg. interception in the case of IHACRES - Jakeman and Hornberger, 1993, cf. Savenije, 2004) it is unlikely that such models could be considered adequate in the physical science sense, regardless of their goodness of fit.

4.5.3 The role of data errors

Data errors are ubiquitous in hydrology and can confound the results of hydrologic studies. For example, for the data used in this study, the streamflow data are subject to uncertainty in the stage-discharge relationship (McMillan et al., 2010), while the
gridded rainfall data are subject to measurement error in the underlying point rainfall data (eg. Nešpor and Ševrůk, 1999) plus interpolation error associated with creating a spatial grid of values based on point measurements (Jones et al., 2009; Tozer et al., 2012).

Although optimisation to a single performance measure (eg. KGE or NSE) remains common in practice, during optimisation the mathematical compensation for input and output errors can lead to spurious results (Thyger et al., 2009). The mathematically optimum parameter set is actually a function of the input and output errors, and a different set of errors may result in an entirely different ‘optimum’ set. In this chapter, since the input and output errors were not explicitly accounted for, the Pareto Fronts generated are similarly a function of the errors in the input and output data. The complex interactions of model structural error with input and output error further complicate the situation (Renard et al., 2010).

The uncertainty in model inputs and flow data propagate through to uncertainty in parameters and projections, and this can be quantified in various ways (eg. Beven and Binley, 1992; Freer et al., 1996; Kavetski et al., 2006a,b; Renard et al., 2010, 2011). Common methods identify not a single parameter set (as in optimisation) but an ensemble of parameter sets, which together are consistent with knowledge of input and output uncertainty, and allow quantification of uncertainty through consideration of multiple possible model simulations.

We affirm that the quantification of uncertainty is an important aspect of any study aiming to provide model projections or forecasting to inform decision making. In contrast, the aim of this study was to revisit the conclusion that rainfall-runoff models suffer from poor performance if applied in climatic conditions different to those against which they were calibrated. Given that previous studies have used single objective optimisation and the DSST to make conclusions about model validity (Vaze et al., 2010) and parameter stationary (Merz et al., 2011), we tailored our method to specifically investigate how reliable the outcomes of such tests may be. The Pareto approach proved useful in this context, but we reiterate that the method used here is not recommended as a general calibration method, in part due to its
4.5. DISCUSSION

inability to estimate predictive uncertainty.

4.5.4 Relevance to future model improvements

The results of this study are instructive towards future efforts to improve rainfall-runoff models. The key lesson for model improvements is this: where improvements are trialled, it is possible that their full benefit will not be seen if evaluated using the DSST in isolation, due to the chance of false negatives. This was shown very clearly (see Figure 4.8 and Figure B.6) for the comparison between GR4J (Perrin et al., 2003) and the modified version GR4JMOD by Hughes et al. (2013). Numerous cases were found where the DSST led to a false conclusion of negligible benefit from the changes of Hughes et al. (2013).

Some studies, such as Brigode et al. (2013) demonstrated a DSST using a method (eg. DREAM - Vrugt et al., 2008) that generated an ensemble of parameter sets. Because such ensemble methods inherently provide information about a wider range of parameter sets, they may be more likely to identify sets that demonstrate the true capabilities resulting from a model improvement. However, this depends strongly on details of methodology, with a key choice being whether or not to explicitly represent the uncertainty of inputs and outputs (eg. Renard et al., 2010, 2011) or adopt objective functions that compensate for these errors without representing them explicitly (as adopted by Brigode et al., 2013, cf. Schoups and Vrugt, 2010).

As discussed above, in this study we did not account for data errors, and so instances of apparent model failure may be related to cases of particularly poor data quality. However, we observed tendencies among catchments where failure was common - namely, they tended to be drier, flatter and have more severe droughts (see also Saft et al., 2015) - and these systematic tendencies support the case for research to better simulate flow generation mechanisms in such catchments, as opposed to assuming that all remaining deficiencies are the result of data errors (Brigode et al., 2013).
4.5.5 Minimising false negatives

This chapter has demonstrated that DSST results may provide a false negative impression of the capabilities of a model. Geometrically, this is associated with Pareto Fronts that had an “inverted L” shape, intersecting regions of robust performance (eg. shaded regions of Figure 4.9), but with endpoint(s) distant from these regions (eg. IHACRES in Figure 4.9). Shapes less prone to false negatives included Pareto curves that formed quasi-linear diagonal lines in the objective space (eg. GR4J in Figure 4.9) and the ideal case (in the sense of parameter stationarity) where the Pareto curve is so compact as to appear as a dot in the objective space (eg. GR4JMOD in 410057, Figure B.2).

It is difficult to generalise about the relation between model complexity (number of parameters) and the tendency to produce false negatives. In a separate analysis (not shown) we examined the shape of the Pareto curves on a model-by-model basis, which demonstrated that the parsimonious model GR4J tended to produce Pareto Fronts of the ‘quasi linear diagonal’ type, and thus have a lower tendency to generate false negative impressions of model capabilities. However, higher model complexity did not necessarily lead to more false negatives, as shown by a comparison of IHACRES (8 parameters, 37 false negatives for Standard 1) and SACRAMENTO (13 free parameters, 32 false negatives for Standard 1).

It is also possible that careful selection of objective functions may minimise false negatives. In the ideal case listed above (Pareto curve collapsed to a dot in the Objective Space), the parameter set identified as optimal in one set of climatic conditions is optimal or near-optimal in other climatic conditions - a desirable attribute for an objective function and/or model structure. In the present context, the tendency to produce this ideal case could be evaluated for a given objective function either by (a) assessing only the endpoints of the Pareto curves (one-at-a-time single objective optimisation, cf. Coron et al., 2012, 2014); or (b) via full Pareto analysis as shown in this chapter. Future research could conduct this analysis individually for a number of objective functions from the literature in turn, and then compare the results. It is
likely that a more nuanced objective function such as a meta-function incorporating responses over multiple timescales (Hartmann and Bárdossy, 2005; Shamir et al., 2005a) may have more success than commonly used functions that consider only the daily timestep. Such analysis would be relevant to the discussion of the value (or lack of value) of single objective optimisation in hydrology (eg. Gupta et al., 2008).

4.5.6 Climate change: beyond the scope of historical observations?

While climate change may be outside of the range of current observations in many regions of the world, in South East Australia the changes in streamflow projected in some climate change studies are of a similar order to the historic streamflow declines during the Millennium Drought. For example, Chiew et al. (2009) projected future runoff across South East Australia using the outputs of 15 Global Climate Models (GCMs). Although there was a high degree of uncertainty, in most locations and for all GCMs the percentage change in long-term average annual flows was generally less than 55% (ibid. Figure 9), which was the median observed reduction during the Millennium Drought for the catchments in this study (Figure 4.4). However, Chiew et al. (2009) used GCM runs based on a 0.9°C increase in temperature, and scenarios with greater temperature increases would result in greater reductions in streamflow that may be beyond the range of observations. Nonetheless, we suggest that it is reasonable to assume that the observed behaviour of catchments during historic dry periods like the Millennium Drought can be used to inform our understanding of possible future behaviour of these catchments under climate change.

4.5.7 Research challenges

In this section we summarise the research challenges to improve rainfall-runoff modelling under a changing climate. These are not original ideas; rather, we aim to relate the present study to existing ideas and trends in the literature. We broadly group the challenges under two headings:

a Making better use of information content of measured data: Figure 4.12 (see
also eg. Oudin et al., 2006; Gupta et al., 2009; Berthet et al., 2010; Andréassian et al., 2012) demonstrated that the use of global performance measures can mask significant deficiencies in simulations. Hydrologists should therefore favour measures that consider a breadth of characteristics about the historic data. The multi-timescale objective functions mentioned above (Section 4.5.5) are an example of this. We also note developments in using hydrologic signatures to inform calibration (Wagener and Montanari, 2011; Vrugt and Sadegh, 2013). While signatures do not inherently take data errors into account, some signatures are less sensitive to data errors than others (Westerberg and McMillan, 2015), so that signature sets can be chosen with the intent of reducing the confounding effect of data errors (Vrugt and Sadegh, 2013) while maximising the information content gained from observed data. This chapter has demonstrated that some existing model structures were capable of simulations that provided robust performance before and after a change in climate. The challenge is to develop calibration methods that can identify these parameter sets using only ‘pre-change’ data. The Pareto method used here does not do this, and furthermore is not viable if the changed climate has not yet been observed.

b Improving process understanding in catchments under change: Following the same logic as above, it may be that even with considerable advances in parameter estimation methods, it is still not possible to identify robust parameter sets using only ‘pre-change’ data. Further research is needed to investigate the physical reasons why runoff is more sensitive to changes in rainfall than current rainfall-runoff models would suggest. Such research would be consistent with the current research focus on change in hydrology and society (IAHS Panta Rhei decade 2013–2022 - Montanari et al., 2013). This knowledge could inform new rainfall-runoff model(s) that, when calibrated to “pre-change” data, would ideally provide more certainty about the trajectory of runoff after a change in rainfall, and be closer to “adequate” in both a physical science and engineering sense. However, it is noted that even if a model does have the correct struc-
ture to simulate flows under contrasting conditions, the relevant parameters may remain poorly identified during calibration (Reichert and Omlin, 1997), depending on the input data and method of calibration.

4.6 Conclusions and recommendations

In this chapter, five conceptual rainfall-runoff model structures were tested in 86 catchments, initially using a Differential Split Sample Test (DSST) that was intended to replicate common practice. When optimized to match the Kling-Gupta efficiency over the nondry period, the models generally had poor performance during the dry period, and vice versa. These results were consistent with existing literature (eg. Vaze et al., 2010; Coron et al., 2012, 2014; Thiel et al., 2015a). Therefore, the model structures largely failed the DSST, although this was catchment dependent. The model structures were then further tested using a Pareto approach via the AMALGAM algorithm. The AMALGAM results demonstrated that many of the cases of apparent failure under the DSST were false negatives. Of the 282 cases of apparent model failure under the DSST using the lower modelling standard (KGE > 0.7), 155 were false negatives. The higher standard (KGE > 0.8) resulted in 347 cases of apparent model failure, with 120 false negatives. Thus, regardless of the standard used, the DSST often missed potentially promising parameter sets within a given model structure.

These results can be used to answer the research question and hypothesis stated at the beginning of the chapter. Responding to the recorded deficiencies of rainfall-runoff model performance in the literature, the research question was: Are current conceptual rainfall-runoff model structures deficient in their ability to simulate streamflow responses to long term changes in climate? The hypothesis to be tested was that the observed poor performance is due to poor or insufficient model calibration and evaluation techniques rather than deficient model structures. The results indicate that this hypothesis was true in around 55% of the cases (155 out of 282) or around 35% of the cases (123 out of 349) of poor performance in
CHAPTER 4. REVISITING MODEL DEFICIENCY IN CHANGING CLIMATE

the DSST, depending on the modelling standard adopted. Thus, the answer to the research question is that some rainfall-runoff model structures are deficient in some catchments, with the corollary that the deficiency is significantly less common than the Differential Split Sample Test might suggest. It was discussed that the definitions of 'deficient' and 'adequate' are themselves dependent on philosophical perspective (Gupta et al., 2012).

As noted throughout the chapter, we are not proposing that the multi-objective approach trialled here is a viable alternative approach to the DSST. The logic expounded by Klemeš (1983) is valid and we affirm the need to withhold a portion of historic data for independent testing and evaluation. The multi-objective approach here does not do this, so the findings of this chapter are based solely on calibration results, with no independent evaluation period. The Pareto approach trialled here is only useful insofar as it has demonstrated that commonly used model calibration and evaluation methods can give a false negative impression of the ability of a model to match observed streamflow.

Based on our results and discussion, we recommend:

a. Caution when interpreting split sample results. Split sample testing remains an essential test of models that will be used operationally (in the sense of Klemeš (1986) and a useful 'first test' of a model structure's capabilities. However, this chapter has demonstrated that split sample test results can give a false negative impression of the ability of a model to match observed streamflow, and are thus a poor basis to reject a model hypothesis.

b. Further work towards identifying parameter sets that are robust to changes in climate. This chapter has demonstrated that commonly used calibration and evaluation methods often fail to identify parameter sets that can simulate flows robustly when climatic conditions change, even when such parameter sets do exist within a model structure. New methods are needed that can more reliably identify such parameter sets.

c. Further research aimed at understanding the physical processes occurring in
catchments when climatic conditions change, in line with the IAHS Panta Rhei
Decade’s focus on change in hydrology and society (Montanari et al., 2013).
Chapter 5

Improving calibration methods, 1:
Exploratory Analysis

5.1 Chapter Summary

1 Rainfall-runoff models are widely used to simulate the impact of changes in climate on water availability. However, studies have shown that rainfall-runoff models often provide poor simulations with high bias when applied in changing climatic conditions. This suggests model structures and/or calibration methods may need to be improved, and this study focussed on the latter. We aimed to identify calibration metrics that can find parameter sets that provide robust simulations across changes in climate. Across a variety of study catchments and model structures, we followed a three-step process of: randomly generating a large ensemble of parameter sets; identifying parameter sets in the ensemble that provided robust simulations both before and after a change (drying) in climatic conditions; and calculating multiple performance metrics for each member of the ensemble to determine which metrics could be used to identify the robust parameter sets using pre-change data only. The performance metrics trialled included traditional objective functions along with less common indices such as the degree of replication of hydrologic signatures. The re-

---

1This chapter has been submitted to the journal Water Resources Research and is currently under revision, as described in the preface.
results indicated considerable scope for improvement of calibration methods, relative to commonly used approaches. Metrics that consider dynamics over a variety of timescales (eg. annual, not just daily) were relatively more promising, as were objective functions using the sum of absolute errors rather than the sum of squared errors. Several promising hydrologic signatures were identified, but the effectiveness of a given signature often varied between model structures. Further tests of shortlisted metrics are undertaken in the next Chapter.

The key points of this chapter are:

- Hydrologic simulations in changing climate strongly depend on choice of calibration metric.

- Common ‘least-squares’ metrics lead to poor split sample results, relative to other metrics.

- Incorporating hydrologic signatures in calibration functions may help, but results suggest this is model-specific.

### 5.2 Introduction

Rainfall-runoff models have potential to be useful tools in planning for future climate variability. They are often used when translating projected climatic shifts (eg. in rainfall or temperature) into projected changes in water availability (eg. Bergström et al., 2001; Chiew and McMahon, 2002; Christensen et al., 2004; Fowler et al., 2007; Chiew et al., 2009; Vaze et al., 2011; Bosshard et al., 2013; Faramarzi et al., 2013; Hagemann et al., 2013; Forzieri et al., 2014). Their utility is particularly important in regions where future projections lie beyond the scope of historical conditions, and/or where available data captures a relatively small portion of historic variability (cf. Gallant et al., 2011; Cook et al., 2016). Applying a rainfall-runoff model in such circumstances assumes that the model remains fit for purpose despite being applied under a changed climatic regime compared to the calibration data (cf. Clarke, 2007; Vaze et al., 2010; Ehret et al., 2014).
5.2. INTRODUCTION

However, existing literature suggests rainfall-runoff models often provide poor simulations with high bias when applied in changed climatic conditions. For example, Vaze et al. (2010) tested four conceptual rainfall-runoff model structures in 61 catchments in South East Australia and noted a reduction in model performance and increase in bias, when evaluated over periods with changed long term rainfall relative to the calibration period, particularly in the case of a transition from wetter to drier conditions. Similar results have been demonstrated for Australian catchments in subsequent studies (eg. Coron et al., 2012; Saft et al., 2016a). Coron et al. (2012) tested 3 conceptual model structures over 216 Australian catchments and reported that “calibration over a wetter (drier) climate than the validation climate leads to an overestimation (underestimation) of the mean simulated runoff” (ibid. p1). These findings are not limited to single regions or model types, with similar results in Europe (Merz et al., 2011; Coron et al., 2014), Africa (Refsgaard, 1997) and North America (Singh et al., 2011); and evidence for similar problems with physically based models (Refsgaard, 1997) and simple annual relationships between rainfall and runoff (Saft et al., 2015).

These problems have prompted a variety of responses from hydrologists. Li et al. (2012) recommend to minimise the degree of extrapolation in climate, stating that “if a hydrological model is set up to simulate runoff for a wet climate scenario then it should be calibrated on a wet segment of the historic record, and similarly a dry segment should be used for a dry climate scenario.” (ibid pl 239; see also Broderick et al. 2016). Similarly, authors such as Singh et al. (2011) and Merz et al. (2011) recommended limits of acceptable change in climatic conditions within which models are more likely to provide acceptable results. Some authors investigated patterns between parameter values and climatic conditions, such as Merz et al. (2011) who conducted separate calibrations of the HBV model on multiple five year segments of the climatic record and noted that the parameters varied systematically with climate, particularly those governing snow melt and the nonlinearity of runoff generation. Many researchers affirm the need for model improvements to better simulate runoff under changing climatic conditions (eg. Merz et al., 2011; Petheram et al.,
2011; Coron et al., 2014) and some studies have produced new model structures in this vein (eg. Ramachurn, 2012; Hughes et al., 2013). The issue of diagnosing and improving nonstationary models is discussed further later in the thesis, in Chapter 7.

A complementary approach is to focus on improving calibration methods (Brigode et al., 2013; Coron et al., 2014; Thiél et al., 2015a). There is growing evidence that common calibration methods, such as optimisation to the Nash-Sutcliffe Efficiency (NSE, Nash and Sutcliffe, 1970) or Kling Gupta Efficiency (KGE, Gupta et al., 2009), often do not choose parameter sets that are robust to changes in climate, even when such parameter sets are available within a model structure, as demonstrated in the previous chapter. In this study, ‘robust’ simulations are those that perform well in split sample testing (Klemeš, 1986) when evaluated over long dry periods that are not part of the calibration period.

The literature contains many examples of alternative calibration methods, some of which could be helpful in changing climatic conditions. Applying transformations prior to least-squares-type calculations may stabilise error variance and thus improve parameter inference (Engeland et al., 2005). As already mentioned, transformations may also place more emphasis on mid-to-low flow (Chiew et al., 1993, 1995; Freer et al., 1996; de Vos et al., 2010; Pushpalatha et al., 2012). Other authors suggest a more balanced consideration of ‘average’ model performance can be attained through absolute-error, rather than squared-error, approaches (Willmott, 1982; Willmott and Matsuura, 2005). Taking a Bayesian statistical approach, Smith et al. (2015) suggested multiple alternative likelihood functions to the common least-squares functions, plus a systematic workflow for choosing between them based on the error characteristics in the data, but their discussion is not specific to changing climatic conditions. Hartmann and Báródy (2005) and Shamir et al. (2005a) demonstrated that robustness could be improved by evaluating catchment response on a variety of timescales (eg. annual) rather than using objective functions/signatures formulated on the daily timestep only. For example, this might cause calibration to consider whether year-to-year variability is matched, in addition
to day-to-day variability. Bärdossy and Singh (2008) investigated the utility of data depth Tukey (1975), a geometric property among parameter sets in an ensemble. Parameter sets with high depth were relatively less sensitive to data errors and more transferable to different time periods. Zhang et al. (2008) among others demonstrate the use of meta-objective functions which consider various aspects of the flow regime separately, then combine the results together into a single ‘meta’ objective function, and this extra information may improve parameter inference. Ghamari et al. (2013) extend this logic to include multiple sub-periods within the calibration period, using a Pareto-based approach to search for parameter sets that provided the best overall compromise over all periods and objectives considered (see also ‘limits of acceptability’ approaches for similar logic applied on a timestep-by-timestep basis - Beven, 2006; Liu et al., 2009). In split sample evaluation, many of these methods outperformed least-squares-based calibration methods, suggesting that the problems outlined above are at least partly associated with calibration methods, and that the numerically optimum parameter set according to least squares measures is often different to the hydrological optimum, defined by Andréassian et al., (2012, p2206) as the parameter set(s) that would ‘work sufficiently outside the calibration period’ - a conclusion supported by the previous chapter.

The present study follows on directly from Chapter 4. Although the method in Chapter 4 identified robust parameter sets with high KGE over both a pre-change (wetter) calibration period and a post change (drier) evaluation period, it used the information in the drier evaluation period to do so, and thus (intentionally) broke the accepted rules of Differential Split Sample Testing (Klemes, 1986). In contrast, the present chapter aims to identify these parameter sets using the information in the pre-change (wetter) period only. It is uncertain whether there is sufficient information content in this period to infer suitable parameter values for a different climatic period, and this will be discussed throughout.

The present study focusses on the case of projected drying (ie. where projections are required for a period that is drier than the calibration period) rather than projected wetting, for three reasons. Firstly, GCM outputs indicate that projected
drying is expected in the study area (southern Australia, cf. Chiew et al., 2009) and is thus a more relevant case study than projected wetting, for this region. Secondly, projections indicate that many other regions around the world may similarly suffer from intensifying water shortage problems, potentially leading to economic hardship and conflict (World Bank, 2016). Improved projections under projected drying are important for planning purposes to mitigate these impacts. Thirdly, the study area provides a case study of an (albeit non-permanent) change from wetter to drier climatic conditions in the form of the thirteen-year-long Millennium Drought, which will be discussed further herein (but see also van Dijk et al., 2013; Aghakouchak et al., 2014; Chiew et al., 2014).

The part of the research is presented in two parts spread over two chapters. This chapter is focussed on a broad exploratory analysis designed to test a wide range of possible objective functions for their efficacy under changing climate. Having developed a shortlist in this chapter, work described in the next chapter aimed to test the shortlisted objective functions in a more rigorous manner.

The aim of this Chapter was:

To search for ways to improve model calibration methods for changing climatic conditions, by analysing patterns relating metric values over a pre-change (wetter) calibration period and model performance over a post-change (drier) evaluation period.

This chapter and the next each have the same hypothesis:

That there exists at least one metric with the property that optimising that metric over a pre-change (wetter) period leads to improved model performance over a subsequent post change (drier) period, compared to the reference calibration method.

The aim and hypothesis are subject to the following definitions. Metric means either an objective function or a hydrologic signature. In the latter case optimising means minimising the error in signature value. Performance refers to the value of the Kling Gupta Efficiency (KGE - Gupta et al., 2009). The KGE was chosen with consideration to the water resources context of this thesis (Chapter 1). For water resource studies, it is important that models can replicate runoff volumes,
at timescales relevant to the water supply system response (days-weeks), and the
three components of the KGE (mean, variability and correlation) are relevant to this
goal. We discuss the implications of using a single metric as our reference metric,
in Section 5.5.4, and other measures of performance are also considered in the next
chapter. The reference calibration method used to represent common practice was
single objective optimisation of the KGE. Although the word subsequent is used
here, this chapter also considered case studies where a drier period was followed by a
wetter period, as per the period definition of the previous chapter (cf. Section 4.3.6).

Subject to these definitions, if the hypothesis is true, then this analysis would
result in at least one calibration method that performs better than the reference
method when evaluated over periods that are drier than the calibration data. The
overall aim of the two chapters together was to identify such a method.

5.3 Method

This section is structured as follows. Section 5.3.1 contains a step-by-step description
of the method. Interpretation is aided by the inclusion of an example case study. Two
broad categories of calibration metric are treated separately: hydrologic signatures
(Section 5.3.1) and objective functions, eg. NSE, KGE etc. (Section 5.3.1).

Subsequent subsections deal with specific aspects of the method, including
definition of signatures and objective functions (Section 5.3.2); models, catchments
and data (Section 5.3.3); rules for selection of case studies for inclusion in the results
(Section 5.3.4); and variations to the method in response to poor sampling density
(Section 5.3.5).

Note that the Kling Gupta Efficiency (KGE, Gupta et al., 2009) is retained as
reference metric (ie., to quantify model performance) even when testing calibration
methods that are not based on the KGE. We chose the KGE over the more commonly
used NSE due to the tendency of the NSE to mask high bias (Gupta et al., 2009, see
also Figure B.4).
5.3.1 Method overview

Introductory case study

To explain the method, consider an example application of the GR4J model structure \cite{Perrin2001} applied in the 104 km\(^2\) catchment of the Stanley River upstream of Peachester, Queensland, Australia (gauge 143303A; mean annual rainfall 1633 mm year\(^{-1}\); mean annual runoff ratio 0.45). The catchment is subject to high annual variability; for example, during the period 1976 to 1982 inclusive, the mean annual runoff was 43\% of the historic average. This was the lowest flow of any consecutive 7-year period on record. Chapter 4 demonstrated that this case study exemplifies the problems described in the introduction, calibrating the GR4J model in this catchment in two separate tests: Test 1 optimised KGE over the 1976 to 1982 period, termed the ‘dry’ period, only; then evaluated the model over the remaining years, termed the ‘nondry’ period (in this case 1951-1975 combined with 1983-2009); and Test 2 optimised KGE over the nondry period and evaluated over the dry period. In each case the GR4J model performed better in calibration than evaluation:

- Test 1: \(\text{KGE}_{\text{dry}}\) (calibration) = 0.94, \(\text{KGE}_{\text{nondry}}\) (evaluation) = 0.71;
- Test 2: \(\text{KGE}_{\text{nondry}}\) (calibration) = 0.86, \(\text{KGE}_{\text{dry}}\) (evaluation) = 0.47;

Test 1 revealed that a good KGE score was possible over the dry period, but Test 2 nonetheless chose a parameter set with a poor KGE score over the dry period. In other words, in Test 2 the calibration method (optimisation to KGE) chose a parameter set that was not robust to a change from wetter to drier conditions. This is consistent with the references cited in the introduction \cite[eg. Vaze et al., 2010; Coron et al., 2012]{Vaze2010} that reported difficulties simulating after a transition from wetter to drier conditions. The question of interest is, can an alternative calibration method be devised, one that selects a more robust parameter set?

To explore this question, we implemented a method based on analysis of patterns in a large, randomly generated ensemble, as outlined in the steps below. The
reason we used random ensembles and not guided approaches was suitability to explore multiple different metrics (objective functions and hydrologic signatures - cf. Section 5.3.2). Whereas a guided approach (for example, a Shuffled Complex Evolution run - Duan et al., 1992) proceeds differently for each objective function and thus would require a separate run for each objective function under consideration, we instead generated a single large random sample of parameter sets and, for each ensemble member, calculated the value of every objective function and hydrologic signature. Avoiding the need for separate calibration runs for each metric thus allowed consideration of a larger list of metrics. The steps below were intended to reduce this list down to a shortlist, prior to testing of shortlisted metrics using guided search approaches in the next chapter.

Method for hydrologic signatures

To explore the question with respect to signatures, we implemented the following steps. These steps were applied to a total of 111 case studies (combinations of catchment and model structure - Section 5.3.4 and 5.3.5). In each case study, 35 individual hydrologic signatures (Section 5.3.2) were considered. Thus, the steps specific to a signature (Steps 4, 9 and 10) were repeated a total of 3885 times.

1. **Generate a large random sample of parameter sets using Latin Hypercube Sampling.** A random sample of at least 100,000 parameter sets was generated using Latin Hypercube Sampling (LHS). In some cases, the sample size was increased to compensate for poor sampling density in interesting regions of the objective space (cf. steps 6 and 8; see also Section 5.3.5).

2. **Run the hydrologic model with each parameter set.** In the example application, this involved 100,000 GR4J simulations of 59 years at a daily timestep.

3. **Plot each parameter set in a 2D objective space based on model performance over two periods with contrasting climate.** Specifically, the x and y axes were formed by: (a) the KGE value over the driest consecutive
CHAPTER 5. IMPROVING CALIBRATION: EXPLORATORY ANALYSIS

Figure 5.1: Illustrative explanation of the method for hydrologic signatures, using the case study of model structure GR4J in Catchment 143303A (Stanley River at Peachester). Scatter plot axes are defined in objective space; parameter sets that are robust to climate change appear in top right. (a) Pareto Curve from Chapter 4 accompanied by 100,000 randomly generated parameter sets, in blue. (b) Shading of parameter sets by the absolute value of a selected hydrologic signature - coefficient of variation, calculated on an annual timestep. (c) Characterising the difference in simulated signatures between the two regions using the Kolmogorov-Smirnov Distance (KSD). The reference calibration method is optimisation of the KGE over the nondry period. In this case the signature is better replicated in the Robust Region (black) than the Reference Method region (grey), meaning that this signature has potential to be used to provide simulations that are more robust in changing climate than those provided by the reference method.

7-year period (KGE$_{dry}$); and (b) the KGE value over the remainder of the timeseries, that is, the years before and after the dry period (KGE$_{nondry}$). These are the same axes used in Chapter 4. Figure 5.1a shows this for GR4J in catchment 143303A. Parameter sets that are more robust to changes in climate will tend to have high values in both x and y; that is, they will be in the top right corner.

4. **Calculate the simulated value of the hydrologic signature of interest, for each parameter set.** 35 different hydrologic signatures were tested, as per Section 5.3.2. For illustration, the random sample in Figure 5.1b is shaded...
by the value of a signature called the $C_v$ annual, calculated from the GR4J simulations. $C_v$ annual is calculated by aggregating streamflow into annual values, then calculating the coefficient of variation ($C_v$) of the annual values (Bárdossy, 2007).

We pause the stepwise description for the following clarifications and observations:

- Throughout this study, all signatures (including those in step 4) were calculated over the nondry (calibration) period only, because the aim is to use the information in the calibration (nondry) period to choose a parameter set suitable in a change from wetter to drier conditions.

- For illustration, we have deliberately chosen an example signature that shows interesting structure. Higher values occupy a broad linear band oriented from lower left to upper right in Figure 5.1b. Values in this band attain a closer match to the observed signature value (0.68).

- This example shows that the replication of $C_v$ annual appears to be associated with robust model performance over changing climates. Consider the parameter sets close to the Pareto Front: the part of the front exhibiting the closest match with the $C_v$ annual is also associated with higher KGE$_{dry}$. Thus, matching the $C_v$ annual over the calibration (nondry) period may lead to improved model performance over the evaluation (dry) period, for this example application.

- However, in Figure 5.1b some of the parameter sets that match the signature value have very poor performance (eg. bottom left corner). Thus, some additional method would be required to screen out these poor parameter sets. This is discussed further in Section 5.5.2 and is the concern of Chapter 6. For Chapter 5, we merely note when promising patterns exist.

The purpose of the remaining steps was to characterise these patterns concisely,
to facilitate comparison across a large sample of catchments. This was done by contrasting the parameter values in two ‘regions’ of the objective space.

5. Calibrate the model over the wetter of the two periods (the nondry period), using the reference calibration method (optimisation of KGE).

This step is Test 2 from the previous section. The resulting parameter set is marked in Figure 5.1b for GR4J in 143303A. The Pareto Curve derived from Chapter 4 is also shown in Figure 5.1, in red. The result of this step is at the left-hand end of the Pareto Curve (by definition).

6. Define a region around the Reference Method parameter set (Euclidean distance ≤ 0.1) shown as a circle in Figure 5.1b. Note that the radius was sometimes increased to compensate for poor sampling density in the region, as discussed in Section 5.3.5.

7. Identify the parameter set closest to the ‘perfect’ point (1, 1). Although it is recognised that the KGE is not a ‘perfect’ descriptor of model performance, we nonetheless adopt this parameter set as an approximation to the ‘hydrologically optimum’ parameter set - the parameter set(s) that would ‘work sufficiently outside the calibration period’ - Andréassian et al. (2012), p2206. This parameter set provides the best trade-off between KGE nondry and KGE dry, weighting both equally.

8. Define the ‘Robust Region’ around the parameter set from Step 7 (Euclidian distance ≤ 0.1) shown as the second circle in Figure 5.1b. As per Step 6, a larger radius was sometimes used (Section 5.3.5).

9. Construct Empirical Cumulative Distribution Functions (ECDFs) of signature values for both regions and define the Kolmogorov-Smirnov D (KSD). The KSD is a non-parametric descriptor of the difference in signature value between the populations in the two regions (Figure 5.1d). This step was subject to a minimum sample size in both regions (Section 5.3.5).
10. **Assign the KSD value a positive sign if the Robust Region provides a closer signature match to the observed value, on average, than the Reference Method region.** Assign a negative sign otherwise. Positive values mean that this signature may be useful in seeking out more robust parameter sets (as in this example). Thus, the results section of this chapter is based primarily on reporting this ‘adjusted KSD’ value.

**Method for objective functions**

The method for objective functions is similar in most respects to that for signatures, so the description below is much abridged. In each of 111 case studies (Section 5.3.5), 15 objective functions (Section 5.3.2) were considered. Thus, the steps specific to an objective function (Steps 4, 9, 10 and 11) were repeated a total of 1665 times.

Steps 1 to 3 are the same as for hydrologic signatures - see previous subsection.

4. **Calculate the objective function value for each parameter set.** For illustration, the random sample in Figure 5.2 is shaded by two different objective functions, for the same case study as the previous section. To avoid confusion, note that the axes of Figure 5.2 are still based on the reference metric (KGE) even when the shading refers to an objective function that is not the KGE. Figure 5.2a is shaded by an objective function called the Split KGE, which is defined in Section 5.3.2. Figure 5.2b is shaded by the NSE calculated on the fifth root of streamflow (cf. Chiew et al., 1993). The patterns in Figure 5.2 are discussed in Step 11 below.

Steps 5 to 8 are the same as for hydrologic signatures - see previous subsection.

9. **Construct ECDs of objective function values for both regions and define the Kolmogorov-Smirnov D (KSD).** This step is identical to Step 9 in the previous section except that it is done for objective function values rather than the simulated signature values.

10. **Assign the KSD value a positive sign if the Robust Region provides a higher objective function value, on average, than the Reference**
CHAPTER 5. IMPROVING CALIBRATION: EXPLORATORY ANALYSIS

Figure 5.2: Illustrative explanation of the method for objective functions, using the same catchment and model as in Figure 5.1, for Split KGE (left) and NSE calculated on the 5th root of streamflow (right). Both objective functions show gradients across the Pareto Front that increase with better evaluation (dry period) performance, as shown in the Empirical Cumulative Distribution Functions (ECDFs, (c) and (d)). Optimisation to Split KGE (graph a) will lead to a parameter set that has good KGE values over both the calibration (nondry) period and the evaluation (dry) period. However, this is less clear with NSE 5th root (graph b) which has a broad plateau of near-optimal values.

**Method region.** Assign a negative sign otherwise. All objective functions in this study are defined so that larger values mean better model performance, with 1.0 being the perfect score.

11. **Take note of the parameter set with the highest objective function value.** This step is necessary to approximate the likely results that would be obtained by an optimiser (cf. Chapter 6). For example, in Figure 5.2, Split KGE attains the highest value very close to the Robust Region, and is therefore a promising objective function in this case study. The NSE 5th root has a broad plateau of near-optimal parameter sets which includes some with low KGE values over both periods. Hypothetically, if the highest value of
5.3. METHOD

NSE 5th root happened to be one of these lower KGE parameter sets, this would demonstrate that a promising KSD value does not guarantee the desired behaviour in an objective function. Step 11 was not required for signatures because, in this study, signature replication is never optimised and signatures are not used in a standalone fashion (cf. the next Chapter).

In summary, for both signatures and objective functions, the method for Chapter 5 identifies signatures and objective functions whose patterns in the objective space (cf. Figure 5.1a) indicate potential to help to identify parameters sets more suited for simulating under a change in climate from wetter to drier. In the results section, signatures and objective functions that consistently show positive values of adjusted KSD values will be spoken of as ‘promising’ - this choice of wording means that the observed pattern (ie. Figures 5.1b; 5.2a; 5.2b) appears to indicate high utility, but this is yet to be confirmed by a guided calibration method. As mentioned, the purpose of Chapter 5 was to test a large range of metrics and develop a shortlist based on these patterns; rigorous testing using guided calibration methods is deferred to Chapter 6. The next section outlines the full range of signatures and objective functions that were tested in Chapter 5.

5.3.2 Metric selection and definition

Guiding principles

Metrics were selected for testing according to two overall principles. Firstly, metrics were selected (or created) that were considered likely, a priori via consideration of hydrologic theory, to result in more robust simulations relative to ‘least squares’ type measures. Secondly, we aimed to develop a set of metrics representative of those used in the hydrologic literature, regardless of any a priori reason to expect more robust simulations from their use in calibration. The latter goal was particularly applied to hydrologic signatures, and was made possible only because the adopted method is easily scalable (ie., it is not much additional effort to test twenty metrics if we are already testing ten).
CHAPTER 5. IMPROVING CALIBRATION: EXPLORATORY ANALYSIS

Each principle is discussed in more detail below. The eventual list of 35 hydrologic signatures and 15 objective functions are given in Tables 5.2 and 5.1, respectively.

Selection of metrics a priori by consideration of hydrologic theory

Based on the literature review in the previous section, numerous classes of objective function were expected to improve model calibration in changing climate, for example:

- **Objective functions that apply transforms to runoff values prior to least-squares calculations.** Smaller exponents increase emphasis on times of mid- and low-flow, which may contain information that is relevant to a drier climate. Transformations may stabilise the variance of errors, improving parameter inference (Engeland et al., 2005) and reducing sensitivity to gauging error during floods (Berthet et al., 2010). Metrics in this category include NSE-square root, NSE-5th root;

- **Greater sensitivity to model bias through a bias penalty.** This counters the tendency of least squares measures to ignore high bias in cases of high flow variability (Gupta et al., 2009). The NSE-bias is the sole metric in this category, although it could be argued that the KGE also applied a greater penalty to bias than NSE does, when flow variability is high (Gupta et al., 2009);

- **Objective functions that use sum of absolute error, rather than sum of squared error approaches.** Not squaring the errors increases emphasis on times of mid- and low-flow, which may contain information that is relevant to a drier climate, and makes calibration less sensitive to gauging errors during floods (Berthet et al., 2010). Metrics in this category include the Index of Agreement and the Volumetric Efficiency;

- **Metrics that consider different aspects of the flow regime and then combine these into a single objective function.** Each measure considers
different information in the calibration data. Consideration of a wider spread of information may improve process representation and parameter inference. The Zhang objective function and the NSE low flow are in this category. It could be argued that every signature fits into this category, as each measures a different aspect of the flow regime. Signatures that characterise aspects not emphasised by ‘least squares’ metrics include the slow flow index (base flow index), the flow duration curve, and the recession constant;

- **Approaches that consider different timescales**, such as signatures or objective functions defined on annual or monthly timesteps, or alternatively time-based meta-objective functions explicitly considering different sub-periods of the calibration period. These may ensure greater attention is given to dry years in the calibration period, which may be a more suitable analogue for drier climate. Such years may be largely ignored by least-squares. Examples of signatures in this category include the dry year fraction, wetting up index, spell counts and durations, $C_\alpha$ annual and skew annual. Objective functions in this category include the NSE monthly, NSE annual and the Split KGE.

In general, these methods were expected to work by broadening the focus of the calibration so that there is less focus on high flows and a greater focus on other aspects of the flow regime, including mid-low flows and/or dry years. The exception is the bias penalty class, which simply corrects the tendency of least squares measures to ignore high bias in cases of high flow variability (Gupta et al., 2009).

**Hydrologic signatures: review and selection from literature**

As noted, the second principle of selection was to test a set of metrics that is representative of those that have been used in the hydrologic literature. This is particularly relevant to hydrologic signatures, and in this section we provide a brief review of hydrologic signature studies, that informed the compilation of the set.

Some of the largest ‘catalogues’ of hydrologic signatures were compiled in the early 2000s by studies in ecohydrology. For example, using data from New Zealand,
Clausen and Biggs (2000) investigated 35 metrics which they grouped into four categories of: magnitude; variability; high flow volume; and high flow frequency. In the context of categorising hydrologic change, Olden and Poff (2003) reviewed 171 metrics, including metrics of low flow behaviour. They found some prior studies had adopted overly large sets of indices, resulting in many signatures with information in common. Note that these studies called hydrologic signatures by a different name such as ‘streamflow variables’ (Clausen and Biggs, 2000) and ‘streamflow indices’ (Olden and Poff, 2003).

Studies aiming to improve prediction in ungauged basins commonly used smaller sets of signatures, sometimes only one or two that were considered most relevant to constraining hydrologic models. For example, Bárdossy (2007) used only the mean and variance of annual flow. Yadav et al. (2007) tested 39 signatures for redundancy, retaining only 13 and noting that 3 signatures were particularly suited to prediction in ungauged basins - high pulse court, runoff ratio and the slope of the Flow Duration Curve. In the context of rainfall-runoff modelling, some methods can incorporate more signatures than others. Shamir et al. (2005b) demonstrated a combined process of signature conditioning (to two signatures) followed by model optimisation, while the procedure of Shamir et al. (2005a) used 10 signatures of different timescales, calibrating the model with a stepwise Bayesian approach. Euser et al. (2013) accounted for the shared information between signatures by conducting a Principle Component Analysis among 8 signatures and basing their analysis on whether a model could replicate the first two components with the same parameter set. Studies based on Approximate Bayesian Computation (ABC) tend to use a smaller number of signatures because this method requires a closer replication of signatures. Nott et al. (2012) used the Flow Duration Curve only (see also Westerberg et al., 2011); while Vrugt and Sadegh (2013) and Sadegh and Vrugt (2014) used the runoff ratio, the baseflow index (called the slow flow index hereafter - signatures 26 and 27) and a two-metric description of the flow duration curve.

The final list of signatures is shown in Table 5.2. The list contains 35 hydrologic signatures, but note that one of these (Signature 7 - Flow Duration Curve) is a two-
metric signature, meaning that a slight variation to the method was required, as set out in Appendix C.3. Typical values of hydrologic signatures are provided in the Appendix, Figure C.3.

5.3.3 Rainfall-runoff model structures, study catchments and data

This Chapter used the same rainfall-runoff model structures as Chapter 4, namely GR4J, SIMHYD, IHACRES, GR4JMOD and SACRAMENTO (Table 5.1). In the present context, the use of numerous and varied model structures aids the defensibility of the results, making it possible to assess whether metrics that produce robust simulations for one model structure also work for other model structures.

Catchments were selected from among the 86 catchments used in Chapter 4 (Figure 2.1; Section 4.3.4; Appendix A). As explained in the following section, it was deemed necessary to select subsets of catchments from this wider list for the purposes of reducing the list of metrics (Tables 5.2 and 5.1) to a shortlist. However, it is noted that Chapter 6 (the companion chapter for the current research question) subsequently tested shortlisted methods on all 86 study catchments. Rainfall, PET and streamflow data was adopted from Chapter 4 - please refer to Section 4.3.5.

Table 5.1: Description of the objective functions that were examined

<table>
<thead>
<tr>
<th>Name</th>
<th>Principal Reference</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NSE</td>
<td><em>Nash and Sutcliffe</em> (1970)</td>
<td>Commonly used in literature and practice</td>
</tr>
<tr>
<td>3. KGE</td>
<td><em>Gupta et al.</em> (2009)</td>
<td>NSE tends to preference lower variability and under-penalise bias.</td>
</tr>
<tr>
<td>4. Index of agreement (refined)</td>
<td><em>Willmott et al.</em> (2012)</td>
<td>Based on minimising the sum of absolute errors rather than the sum of squared errors (cf. Legates and McCabe, 1999)</td>
</tr>
<tr>
<td>5. Volumetric efficiency</td>
<td><em>Criss and Winston</em> (2008)</td>
<td>Similar to the previous entry, but a slightly different formulation.</td>
</tr>
<tr>
<td>Name</td>
<td>Principal Reference</td>
<td>Comment</td>
</tr>
<tr>
<td>------</td>
<td>---------------------</td>
<td>---------</td>
</tr>
<tr>
<td>6. NSE with a bias penalty</td>
<td>Vaze et al. (2010)</td>
<td>Penalty function increases non-linearly as the absolute value of bias increases.</td>
</tr>
<tr>
<td>7. NSE on square root of flows (0.5)</td>
<td>Chew et al. (1995)</td>
<td>In certain contexts this has been shown to stabilise heteroscedasticity (eg. Engelund et al., 2005). This provides similar results to using the log but sidesteps problem caused by cease to flow periods (namely, taking the log of zero).</td>
</tr>
<tr>
<td>8. NSE on fifth root of flows (0.2)</td>
<td>Chew et al. (1993)</td>
<td>Same as normal NSE but the reciprocal of the flows is used. Recommended for low flow matching by (Pushpalatha et al., 2012).</td>
</tr>
<tr>
<td>9. NSE$_{15}$</td>
<td>Pushpalatha et al. (2012)</td>
<td>Aggregate daily flows to monthly and then calculate NSE using monthly values.</td>
</tr>
<tr>
<td>10. NSE monthly</td>
<td>Vaze et al. (2011)</td>
<td>Aggregate daily flows to annual and then calculate NSE using annual values.</td>
</tr>
<tr>
<td>11. NSE annual</td>
<td>Hartmann and Bárdossy (2005)</td>
<td>NSE annual and the NSE daily, equally weighted. This metric splits the timeseries into individual years, calculates the KGE value for each year in isolation, and then takes the average.</td>
</tr>
<tr>
<td>12. NSE annual + NSE daily</td>
<td>Hartmann and Bárdossy (2005)</td>
<td>As with the previous, but the metric is the average of the lowest 25% of annual values.</td>
</tr>
<tr>
<td>13. Split KGE</td>
<td>This chapter</td>
<td>Equally weighted combination of metrics regarding high flows, low flows, timing and bias.</td>
</tr>
<tr>
<td>14. Split KGE 25</td>
<td>This chapter</td>
<td></td>
</tr>
</tbody>
</table>

**5.3.4 Rules for inclusion of case studies**

As mentioned, not all the 86 catchments were used for Chapter 5. This section explains why subsets were used and how they were selected. Hereafter in this chapter we use the phrase *case study* to refer to a combination of model structure and catch-
Table 5.2: Description of the hydrologic signatures that were examined

<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean daily</td>
<td>Average daily flow (mm/d). Note: in this context, the mean and runoff ratio deliver identical results. We adopt only the mean. In the context of regionalisation (eg. Yadav et al., 2007) they do not deliver identical results, so it is necessary to specify.</td>
<td>Clausen and Biggs (2000); Olden and Poff (2003); Shamir et al. (2005a); Yadav et al. (2007); Bárdossy (2007); Sadegh et al. (2015)</td>
</tr>
<tr>
<td>2. Median daily</td>
<td>Median daily flow. (mm/d)</td>
<td>Olden and Poff (2003); Sadegh et al. (2015)</td>
</tr>
<tr>
<td>3. Median annual</td>
<td>Median value of annual flow (mm/yr)</td>
<td>This study only</td>
</tr>
</tbody>
</table>

**Flow duration, variability and skewness**

| 4. $C_v$ daily    | Coefficient of variation of daily flow. (-)                                | Clausen and Biggs (2000); Olden and Poff (2003); Yadav et al. (2007); Sadegh et al. (2015) |
| 6. Skew daily     | Skewness of daily flows, calculated here as the Yule-Kendall index (-)     |                                                                               |
| 7. Skew annual    | Skewness of annual flows, calculated here as the Yule-Kendall index (-)     | This study only                                                               |
| 8a. FDC fitting 1 | Two metric description of the daily flow duration curve using the van-Genuchten form used in soil hydraulics. Note: FDC fit 1 values are logged (base 10) relative to Vrugt and Sadegh (2013). | Vrugt and Sadegh (2013); Sadegh and Vrugt (2014); Sadegh et al. (2015) |
| 8b. FDC fitting 2 |                                                                               |                                                                               |
| 9. IQR ratio      | Ratio of the 75th and 25th percentile daily flow. (-)                       | Olden and Poff (2003); Sadegh et al. (2015)                                   |
### Timing and seasonality

<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Dry year fraction</td>
<td>Ratio of the 10th percentile annual flow to the mean annual flow. ((\cdot))</td>
<td>This study only</td>
</tr>
</tbody>
</table>

**11. Constancy of flow**

Measure of the degree to which a state does not change with time, based on the definition of Colwell (1974), using monthly flow and 11 flow classes. (\(\cdot\))

*Clausen and Biggs (2000)*

**12. Predictability of flow**

Measure of the degree to which a state is predictable in time, either due to constancy, or regularity of seasonal patterns.

Based on the definition of Colwell (1974), using monthly flow and 11 flow classes. (\(\cdot\))

*Clausen and Biggs (2000)*

**13. Wetting up index**

Normalised by the mean annual precipitation. (\(\cdot\))

This study only

**14. Wetting up index annual variability**

Year-to-year variability in the value of the wetting up index, expressed as the coefficient of variation (\(\cdot\))

This study only

**15. Wetting up index proportion of years**

Proportion of years where flow was greater than the threshold used in the wetting up index (in this case, 25\% of mean annual flow). (\(\cdot\))

This study only

### Low flows

<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16. Spell count - annual average</td>
<td>Average number of occurrences, per year, during which the flow remains below a certain threshold. After <em>Oden and Poff (2003)</em> we adopt the 25th percentile. (per year)</td>
<td><em>Oden and Poff (2003); Sadegh et al. (2015)</em></td>
</tr>
</tbody>
</table>

**17. Spell count - annual variability**

Year to year variability in the number of low flow spells, expressed as the coefficient of variation. (\(\cdot\))

*Oden and Poff (2003)*
5.3. METHOD

<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18. Spell duration average</td>
<td>Average length of low flow spells. (days)</td>
<td>Olden and Poff (2003); Salegh et al. (2015)</td>
</tr>
<tr>
<td>19. Spell duration variability</td>
<td>Variability in the length of low flow spells, expressed as the coefficient of variation. (-)</td>
<td>Olden and Poff (2003)</td>
</tr>
</tbody>
</table>

**High flows**

| Spell count - annual average | Average number of occurrences, per year, during which the flow remains above a certain threshold. After Olden and Poff (2003) we adopt the 75th percentile. (per year) | Olden and Poff (2003); Salegh et al. (2015) |
| Spell count - annual variability | Year to year variability in the number of high flow spells, expressed as the coefficient of variation. (-) | Olden and Poff (2003) |
| Spell duration - average | Average length of high flow spells (days) | Olden and Poff (2003); Salegh et al. (2015) |
| Spell duration - variability | Variability in the length of high flow spells, expressed as the coefficient of variation. (-) | Olden and Poff (2003) |
| Peak distribution slope | Average slope between the 10th and 50th percentiles of the exceedance curve, where the curve is only composed of local maxima in the timeseries of flow (-) | Euser et al. (2013); Salegh et al. (2015) |
| High flow discharge | 90th percentile flow (mm/d) | Clausen and Biggs (2000); Olden and Poff (2003); Yadav et al. (2007) |

**Rates of change**

<p>| Slow flow index, ( \alpha = 0.925 ) | a.k.a Baseflow index. Slow flow as a proportion of total flow, where slow flow is defined using the Lyne and Hollick filter with ( \alpha = 0.925 ) (Nathan and McMahon, 1990) (-) | Singh et al. (2011); Vrugt and Salegh (2013); Salegh and Vrugt (2014); Salegh et al. (2015) |</p>
<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>27. Slow flow index, $\alpha = 0.980$</td>
<td>Same as previous but with $\alpha = 0.98$ (based on Australian field studies, <em>CSIRO and Sinclair Knight Merz (SKM)</em>, 2010; <em>Ladson et al.</em>, 2013) (-)</td>
<td><em>CSIRO and Sinclair Knight Merz (SKM)</em> (2010); <em>Ladson et al.</em> (2013)</td>
</tr>
<tr>
<td>28. Recession constant</td>
<td>Mean daily flow decline, defined as a percentage, on days when flow is decreasing. (%)</td>
<td><em>Yadav et al.</em> (2007); <em>Sadegh et al.</em> (2015); <em>Oden and Poff</em> (2003)</td>
</tr>
<tr>
<td>29. Recession variability</td>
<td>Variability in the day-to-day values of recession constant, expressed as the coefficient of variation. (-)</td>
<td>This study only</td>
</tr>
<tr>
<td>30. Recession constant (slow flow)</td>
<td>Same as above, except to avoid noise induced by ‘flashy’ behaviour, we redefine this using the timeseries of slow flow ($\alpha = 0.925$, see above). (%)</td>
<td>This study only</td>
</tr>
<tr>
<td>31. Positive flow change magnitude</td>
<td>Mean of difference between natural log-arithmetic of flows between two consecutive days of increasing flow (mm/d, logged)</td>
<td><em>Oden and Poff</em> (2003)</td>
</tr>
<tr>
<td>32. Rising limb density, daily</td>
<td>Number of local maxima divided by the length of time that the hydrograph is rising. (per day)</td>
<td><em>Shamir et al.</em> (2005b); <em>Euser et al.</em> (2013); <em>Sadegh et al.</em> (2015)</td>
</tr>
<tr>
<td>33. Declining limb density, daily</td>
<td>Number of local maxima divided by the length of time that the hydrograph is declining. (per day)</td>
<td><em>Shamir et al.</em> (2005b); <em>Sadegh et al.</em> (2015)</td>
</tr>
<tr>
<td>34. Catchment lag (integer)</td>
<td>Lag between rainfall and resultant streamflow, calculated via correlation analysis.</td>
<td>This study only</td>
</tr>
<tr>
<td>35. An answer of $x$ indicates that a lag of $x$ days results in the highest possible correlation between rainfall and shifted streamflow. (days)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3. METHOD

<table>
<thead>
<tr>
<th>Short name</th>
<th>Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>35. Catchment lag (non integer) 12-15 from Perrin et al. (2001), which have a single parameter (time base in days). An answer of x indicates that a time base of x days results in the highest possible correlation between the shifted rainfall and streamflow. (days)</td>
<td></td>
<td>This study only</td>
</tr>
</tbody>
</table>

ment. That is, a case study is a model structure applied to a catchment. Since there are 86 catchments and 5 model structures, there are a total of 430 case studies that a given signature or objective function could be tested on.

The method explained above (Section 5.3.1) summarised patterns in the objective space via a single number (the adjusted KSD), allowing comparison and synthesis across catchments. In this synthesis, it is logical to focus on catchments where models performed well during wetter periods but poorly during drier periods, since this is the behaviour that prompted this study (Section 5.2). Furthermore, it is useful to ignore case studies where the Pareto analysis of Chapter 4 indicated that no parameter set exists that attains a high KGE value over both nondry and dry periods. Since no parameter sets exist with the desired behaviour, improving simulations for these case studies is beyond the scope of any calibration method, no matter how good. The following two rules were used to select case studies for Chapter 5:

a. Robust performance must be possible over both the dry period and the nondry period, using the same parameter set. Case studies were selected if Chapter 4 classified them as capable of KGE > 0.7 in both climatic periods tested, with the same parameter set (see Section 4.3.6 for climatic period definitions). In other words, case studies were selected if we knew, a priori, that good model performance was possible in both dry and nondry periods using the same parameter set.
The parameter set chosen by the reference method did not fulfil rule 1. The intention was to select case studies where the reference method (optimisation to KGE) failed to identify a parameter set with robust performance, even though such a parameter set did exist as per (1). Chapter 4 used the phrase 'false negative' to describe such a situation. The specific rule applied was: a case study was selected if the reference method chose a parameter set with a $\text{KGE}_{dry}$ value at least 0.2 lower than the $\text{KGE}_{dry}$ value of the 'robust region' parameter set (cf. step 7 in Section 5.3.1).

Taken together, these rules tended to select only those case studies where high KGE performance was possible over the two different climatic conditions tested, but the reference method failed to select a parameter set that provided this robust performance.

These rules resulted in a reduced set of case studies, as per the lists below. Since there are 86 study catchments, each model structure had a possible 86 case studies, but the application of the selection rules reduced the number to between 18 and 31 case studies per model structure:

- **GR4J**: 18 case studies (0 failed both (1) and (2); 35 failed (1) only; 33 failed (2) only);
- **SIMHYD**: 18 case studies (13 failed both (1) and (2); 30 failed (1) only; 25 failed (2) only);
- **IHACRES**: 24 case studies (8 failed both (1) and (2); 11 failed (1) only; 43 failed (2) only);
- **GR4JMOD**: 31 case studies (0 failed both (1) and (2); 25 failed (1) only; 30 failed (2) only); and
- **SACRAMENTO**: 26 case studies (6 failed both (1) and (2); 20 failed (1) only; 34 failed (2) only).

Note that this list was reduced further because of difficulties obtaining sufficient sampling density, as outlined below.
5.3.5 Sampling method

As explained in Section 5.3.1, the method involved comparing the metric values across two regions in the objective space, the Reference Method Region and the Robust Region, each shown as circles in Figure 5.1. As with any comparison between two populations, a large population in each region was preferable as it reduced the probability of spurious results arising from random chance. This section describes the procedure for ensuring, where possible, adequate sampling density in both regions.

As mentioned in Section 5.3.1, all random samples were generated using LHS. Initially, a random sample of 100,000 parameter sets was taken. In some cases, numerous parameter sets (ie. $>100$) fell within each of the two regions, so that there was considerable (albeit subjective) confidence that any apparent pattern was real and was not simply the product of random selection. However, the cases where one or both regional populations was small (ie. $<10$) required more careful consideration. There are many statistical tests available to test for the difference between two samples, but few that are helpful in this context to choose how many samples should be acquired. One possible course would be to continue sampling to increase the sample sizes in the regions until a statistically significant difference is observed between the regions. However, in some cases there could be no difference between the underlying populations of the regions; if so, such a method could result in an arbitrarily large number of samples being produced with no result obtained. If it were known a priori what regional difference should be expected, then it might be possible to infer the number of samples required. However, in the present case nothing was known a priori, so a subjective limit was used instead, as described below.

If the population in either region was less than the subjectively chosen limit of twenty samples, the following measures were taken. The size of the random sample was increased iteratively in intervals of 100,000 samples each, up to a maximum of 3,000,000. If there were still less than twenty samples in either region, the affected region(s) were expanded in the objective space. The shape of the regions was always circular but the radius was increased to incorporate a larger sample of parameter
sets, subject to the limit that the two regions could not overlap. If both these measures were exhausted (i.e. larger sample and larger regions) without both regions incorporating at least twenty parameter sets each, the case study was rejected for inclusion in Chapter 5.

These measures were required as follows:

- GR4J: neither measure: 14, larger sample: 4; larger sample and larger region(s): 0; rejected: 0.
- SIMHYD: neither measure: 16, larger sample: 2; larger sample and larger region(s): 0; rejected: 0.
- IHACRES: neither measure: 1, larger sample: 8, larger sample and larger region(s): 12; rejected: 3.
- GR4JMOD: neither measure: 10, larger sample: 11, larger sample and larger region(s): 7; rejected: 3.
- SACRAMENTO: neither measure: 12, larger sample: 9, larger sample and larger region(s): 5; rejected: 0.

Thus, the total number of case studies was 111. The chosen case studies were relatively well spread geographically; for example Table C.1 in the Appendix shows that each state of Australia that was represented in the set of 86 catchments was also represented in the set chosen for each model. The results shown in the next section are based only on these selected case studies, but note that the all case studies (rejected or not) were included in the final analysis described in Chapter 6.

5.4 Results

5.4.1 Hydrologic signatures

Figure 5.3 shows the KSD values for each of the 35 hydrologic signatures, for each of the 5 model structures. Recall that the KSD values have been adjusted by giving the
value a negative sign if the signature tended to be better matched in the reference method region. Positive values indicate a closer match in the robust region and are therefore the desired result, indicating potential for use in calibration methods for changing climates.

In Figure 5.3, the first and fourth signatures (mean daily and $C_0$ daily) were more consistently negative than most other signatures. This is not surprising because both signatures are components of the KGE (Gupta et al., 2009) and the experimental design implies a decline in KGE in the robust region compared to the reference method region (cf. Figure 5.1a). However, most signatures were more evenly spread between positive and negative values, and some signatures showed a clear bias towards the positive side, indicating potential for use in calibration methods for changing climates. The following subsection discusses promising signatures for each model structure separately.

**Results for individual model structures**

As per Section 5.2 and 5.3, the purpose of this analysis was to develop a shortlist of metrics that showed promise for use in a changing climate. Judging by the median in each boxplot of Figure 5.3, the four most promising signatures for each model structure were:

- **GR4J**: $C_0$ Annual, Peak Distribution Slope, Dry Year Fraction, Wetting Up Annual $C_0$;
- **SIMHYD**: High Flow Discharge, Constancy, Dry Year Fraction, Skew Annual;
- **IHACRES**: High Flow Discharge, Slow Flow ($\alpha = .925$), Wetting Up Annual $C_0$, IQR ratio;
- **GR4JMOD**: $C_0$ Annual, High Spell Duration $C_0$, IQR ratio, Falling Limb Density;
- **SACRAMENTO**: Dry Year Fraction, $C_0$ Annual, Flow Duration Curve metric 1, Skew Annual;
CHAPTER 5. IMPROVING CALIBRATION: EXPLORATORY ANALYSIS

There were some cases where the adjusted KSD value seemed to show promise, but the actual ECDFs indicated otherwise. This was particularly the case if both regions enveloped the observed value but showed a very wide range of signature values. This problem affected the above lists: cases marked with an asterisk (*) indicate that a signature was removed from the list after manual inspection of the ECDFs revealed this problem. The removed signatures were High Spell Duration \( C_v \) (SIMHYD, IHACRES) and High Spell Count (IHACRES, SACRAMENTO). Combined ECDFs for these signatures, showing the above problem, are provided in the Appendix, Figure C.1. In addition to the above issue, the IHACRES list contained both Slow Flow signatures; to maintain diversity in the shortlist only one of these was retained.

It is noteworthy how many signatures in the shortlist are defined, at least partly, on an annual basis, namely \( C_v \) annual (three instances), Dry Year Fraction (three instances), Wetting Up \( C_v \) (two instances), and Skew Annual (two instances). This is discussed further in Section 5.5.1.

Results across all model structures

At first glance, the above shortlist indicates some consistency between models, since certain signatures were chosen by multiple model structures (eg. \( C_v \) annual; Dry Year Fraction). Can this be extended to identify certain signatures which are promising (or at least neutral) for all model structures? Unfortunately, a detailed inspection of Figure 5.3 indicates not - for example:

- Matching the \( C_v \) annual is likely to encourage robust performance for GR4J, GR4JMOD and SACRAMENTO, but discourage robust performance for SIMHYD and IHACRES;

- Matching the Dry Year Fraction is likely to encourage robust performance for all model structures except IHACRES for which it is likely to discourage robust performance;

- Matching the High Flow Discharge is likely to encourage robust performance
Figure 5.3: Boxplots of adjusted KSD values for each hydrologic signature, presented by model structure. Positive values indicate potential for use in calibrating models for changing climates, due to closer signature match in the robust region compared to the reference method region. ‘Closer match’ refers to replication over the non-dry (calibration) period. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and denoted as +. The plot indicates that some hydrologic signatures show promise (positive values) for use in calibration for changing climates, such as Cv annual in GR4J and the High Flow Discharge in IHACRES. However, there is no signature that shows promise across all five model structures tested.
for SIMHYD and IHACRES whereas it is broadly neutral for SACRAMENTO and is likely to discourage robust performance for GR4J and GR4JMOD.

In other words, in the context of producing robust simulations under changing climates, hydrologic signatures that are helpful for one model structure are commonly unhelpful in others. This suggests that it is unlikely (but not impossible) that there is a single set of signatures that can be used in calibration to produce robust simulations regardless of model structure; instead, a successful calibration strategy based on signatures should adopt signatures most relevant to the model structure in question. This is a difficult finding because it implies that the modeller should have a priori knowledge of which signatures will work best with the model structure in question.

### 5.4.2 Objective functions

Figures 5.4 and 5.5 show results for objective functions. Recall from Section 5.3.1, Step 11 that the KSD plots for objective functions (Figure 5.4) must be supplemented by information about the parameter set with the highest calibration value. Figure 5.5 was derived by plotting this parameter set, for each catchment, in the same two-dimensional objective space used in Figures 5.1 and 5.2, then shading the figure according to how frequently the result landed within each square of a coarse grid. Objective functions with high frequency (yellow) in the top right hand corner are desired as they indicate robust simulations.

Figure 5.4 is useful primarily because it is directly comparable with Figure 5.3; this comparison shows that results for objective functions were more consistent across model structures than for hydrologic signatures. In particular, two objective functions based on the absolute error rather than sum of squares - namely, the Index of Agreement and the Volumetric Efficiency - showed promising results across all five model structures. Inspection of the outputs revealed that these two objective functions delivered almost identical parameter sets; to reduce redundancy, only one was shortlisted.

The KGE results require special consideration due to its role as the reference
Figure 5.4: Boxplots of adjusted KSD values for each objective function, presented by model structure. Positive values indicate potential for use in calibrating models for changing climates, due to higher OF score in the robust region compared to the reference method region. "Higher OF score" refers to the objective function score over the nondry (calibration) period. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and denoted as +. This plot indicates that some objective functions show promise (positive values) for use in calibration for changing climates, such as the Refined Index of Agreement ("Ind. Agree").
Figure 5.5: Gridded results for objective functions showing the KGE values in calibration (y axis) and evaluation (x axis) attained by the objective functions for each model structure. The colours show frequency, i.e. how often did the ‘optimum’ parameter set lie in a given grid square, across all catchments? Note that no optimiser was used; rather, the focus is on the highest objective function score out of the random sample. However, the scores themselves were calculated over the nondry period, hence its labelling as the ‘calibration’ period. High frequencies in the top right hand corner indicate that an objective function identifies parameter sets that are robust to changes in climate, using data in the nondry (calibration) period only.
metric. In Figure 5.4, the KGE showed consistently negative values of adjusted KSD. As with signatures mean and $C_v$ (Section 5.4.1), this was expected since the experimental design implies a decline in KGE in the robust region compared to the reference method region (cf. Figure 5.1a). In the grid plots (Figure 5.5) the highest value of KGE in calibration (nondry period) was consistently in the top two rows (>0.6), which is a result of selection rule 1 in Section 5.3.4. Likewise, the KGE in evaluation (dry period) was never in the right-hand column (ie. was always <0.8) due to selection rule 2. Given these selection rules, Figure 5.5 should not be considered a fair comparison of the abilities of KGE with the other objective functions because the selection rules imply a significant disadvantage. This disadvantage was removed in Chapter 6 which tested all shortlisted objective functions in all catchments for all model structures, treating each objective function identically (next chapter).

Nonetheless, the gridded plots in Figure 5.5 show that numerous objective functions found robust parameter sets where the KGE had failed to do so, including the aforementioned Index of Agreement and Volumetric Efficiency. In addition, there were considerable differences between the Split KGE and the standard KGE even though the Split KGE is essentially the same objective function, the only difference being that the latter is scaled so that no year can have more influence than any other year (Table 5.1). Thus, as with signatures, consideration of dynamics on annual timescales, in addition to daily, appeared to yield significant benefits. However, this was not universal - the objective function that equally weighted NSEdaily and NSEannual (“NSE ann-dly”) did not show significant benefits over the NSEdaily, in contrast with the results of Hartmann and Bánossy (2005).

Given the consistency across model structures, a single shortlist of objective functions was selected:

- Index of Agreement (as stated, Volumetric Efficiency was not included, to reduce redundancy);

- NSE on square root of flows and NSE on fifth root of flows, both of which provided advantages over NSE in most case studies, and variants of which are
widely used in the literature;

- Split KGE and Split KGE 25;

- The Zhang combination, which was the only example of a weighted average of contrasting objective functions, and was thus included to maintain diversity in the shortlist.

In addition, the KGE was retained as reference method for Chapter 6 and was thus tested along with the others; likewise, the NSE was tested due to its assumed familiarity with readers.

5.4.3 Results summary

In summary, the exploratory analysis showed that, for hydrologic signatures:

- Numerous signatures were identified for which closer replication was associated with more robust simulations over changing climatic conditions;

- In particular, signatures considering annual timescales showed relatively greater promise;

- However, there was relatively low consistency between models: hydrologic signatures that were helpful for one model structure were commonly unhelpful in others.

The results for objective functions showed:

- There was greater consistency between model structures for objective functions;

- Objective functions based on absolute error rather than the sum of squares were promising; and

- (as with signatures) consideration of annual timescales showed potential to improve simulations.
5.5 Discussion

5.5.1 Possible explanations for successful metrics

This section interprets some common themes noted in the previous section, beginning with objective functions. The two main findings with regards to objective functions, namely (i) the success of sum-of-absolute-error methods; and (ii) the success of the Split KGE methods, can both be explained with similar reasoning. Their success is related to the relative emphasis placed on different days in the calibration data. NSE and KGE sum the square of each error and significantly weight the mean (Gupta et al., 2009). As noted regularly in the literature, this places emphasis on high flow days because data errors are heteroscedastic (Sorooshian and Dracup, 1980) and because floods contribute significantly towards the mean. Large random data errors on relatively few days of high flow may unduly influence parameter selection, eroding simulation reliability. These impediments are reduced by the successful methods, as follows:

- Sum-of-absolute-error methods are still sensitive to high flow days but provide more balanced consideration of medium and low flows because the errors are not squared. This leads to higher weighting given to dry periods within the training data, relative to sum-of-squares methods. Similar logic applies to the transformed versions of the NSE (square root and fifth root). The success of these metrics in conditioning models for changing climates is an interesting addition to existing literature debating the common reliance on 'least squares' measures when evaluating models (eg. Willmott, 1981; Willmott and Matsumura, 2005; Willmott et al., 2009; Chai and Draxler, 2014).

- Split methods such as Split KGE and Split KGE 25 place hard limits on the influence of wet periods. With standard KGE, parameter selection may have been dominated by the wettest few years in the training data, but the Split KGE limits the influence of each year to $1/N$ (where $N =$ number of training years). The Split KGE 25 focusses only upon the years of poorest KGE, so
that poorly simulated drought behaviour is no longer ignored by the objective function.

It is likely that the success of annual-based signatures can be traced to similar origins. For example, focussing on Annual $C_v$ in GR4J, the factors discussed above may cause GR4J to largely ignore the dry training years, possibly leading to simulations that are too high during these years, thus underestimating variability of annual flow (Figure 5.11). Forcing the model to more closely match the Annual $C_v$ limits these tendencies, leading to better simulations when GR4J is subsequently applied in dry conditions. However, it is notable that neither Annual $C_v$ nor any of the other signatures work universally. It seems reasonable that the above logic should apply also to SIMHYD and IHACRES, but the results indicate that Annual $C_v$ does not encourage robust simulation in these models, even though other annual-based signatures do (Dry Year Fraction and the Wetting Up Annual $C_v$, respectively). Given these interesting and rather perplexing results, this is clearly a fertile area for further research.

5.5.2 Value of matching single hydrologic signatures

As shown in Figure 5.1, matching a single signature is not sufficient as a standalone calibration strategy. Commonly, many parameter sets that match a given signature have very poor model simulations as judged by KGE value. How then can the promising patterns identified in this chapter be exploited? As noted in Section 5.3.1, an additional method is required to screen out the poorly performing sets. For example, this might be achieved by:

i. a hybrid method that considers both the KGE$_{nondry}$ value and the match with the signature(s). For example, first identify all parameter sets in a random sample above a certain threshold in KGE$_{nondry}$; then, from within this reduced sample, select the parameter set(s) with closest signature replication; or

ii. a method based purely on hydrologic signatures that uses other hydrologic signatures to screen out the poorly performing parameter sets. For example,
since KGE is so closely related to the mean and the $C_p$, a method that requires not only the replication of the ‘promising’ signature but also of the mean and $C_p$, might be sufficient to screen out many poorly performing sets.

This logic is further discussed in Chapter 6 of this research questions (next chapter).

5.5.3 Random samples versus guided approaches

In considering the efficacy of calibration methods such as (i) and (ii) above, it is prudent to consider the lessons from this chapter regarding sampling density. As explained in Section 5.3.5, a random (LHS) sample of 100,000 parameter sets was insufficient to gather 20 samples in both regions for 64 out of 117 separate case studies, and a sample of 3,000,000 was insufficient in 6 out of 117 cases. We conclude that the volume of the parameter space occupied by robust parameter sets is sometimes very small and may be missed by a random sample, particularly for more highly parameterised models.

In the context of changing climate conditions, this suggests methods that rely on random sampling may not be reliable, such as method (i) above or non-guided variants of GLUE (Beven and Binley, 1992; Beven, 2006, but cf. Vrugt and Beven, 2016). Rather, this context may require guided methods, ie. methods that proceed through the parameter space in the direction of highest performance (eg. Duan et al., 1992; Kuczera and Mroczkowski, 1998; Vrugt and Robinson, 2007; Vrugt et al., 2008). This is not to say that guided techniques are guaranteed to identify robust sets should they exist (cf. Beven and Binley, 2014, p5904) - indeed, this study is premised in the idea that finding such parameter sets is difficult and requires more research. However, the comparison of random sampling results with results from Chapter 4’s Pareto curves generated by the guided technique AMALGAM (Vrugt and Robinson, 2007) show that random sampling missed significant robust regions of the parameter space, as indicated by the ten examples in the Appendix, Figure B.5. Together, these ten plots demonstrate the following principle: just because a random sample has not
found a robust parameter set, does not mean that none exist.

This implies that Figure 5.5 above, which summarises method efficacy based on the best score achieved by a random ensemble, may be unreliable. The next chapter remedies this using a guided search technique, the CMA-ES optimiser Hansen et al. (2003). This provided a comparable but more defensible analysis to that shown in Figure 5.5. With respect to signature matching (eg. method (ii) in the previous section) it is highly unlikely that a random ensemble will contain many parameter sets that closely match multiple signatures at the same time (Vrugt and Sadegh, 2013). This suggests a guided approach to signature matching is needed, and this is discussed further in Chapter 6 with reference to the DREAM-ABC method (Sadegh and Vrugt, 2014).

Notwithstanding the above, for this study the random sampling approach had significant utility: it reduced a wide range of signatures and objective functions down to a shortlist, a task that would have been very computationally intensive using guided approaches because a separate guided run would have been required for each metric under consideration.

### 5.5.4 Kling Gupta Efficiency as a reference metric

The reference metric (Kling Gupta Efficiency or KGE, Gupta et al., 2009), like any global measure, does not give a complete picture of model performance - it may mask periods of poor performance (eg., see Figure 4.12 from the previous chapter) and ignores aspects of the flow regime that are not strongly related to the mean, the variability or the correlation. Furthermore, a low score in KGE is difficult to interpret on its own (Gupta et al., 2008): is it due to bias, over/underestimation of variability, poor correlation, or a mixture?

It is important to use a wide variety of metrics to evaluate the quality of simulations in hydrology (Gupta et al., 2008), particularly in the context of changing climatic conditions (Thirel et al., 2015b). However, given the number of metrics and case studies considered in this chapter, we assume - for Chapter 5 only - that the KGE is sufficient to judge between metrics, deferring consideration of other
evaluation metrics to Chapter 6. Chapter 6 considers a smaller number of signatures and objective functions in greater detail, allowing more scope for exploring different aspects of the flow regime.

5.5.5 Recommendations

Based on Chapter 5 results, we recommend:

- The hydrologic signatures and objective functions identified in this chapter should be further tested using guided search methods as described in Section 5.5.3, and this is done in Chapter 6. Specifically:
  
  - For objective functions: Index of Agreement; NSE$_{square-root}$; NSE$_{fifth-root}$; Split KGE, Split KGE 25 and Zhang Combination [testing on all model structures]. It is noted that the Volumetric Efficiency was found to deliver similar results to the Index of Agreement, and was not included in the shortlist to reduce redundancy.

  - For hydrologic signatures: that the following signatures be tested, only for the model(s) for which they were associated with robust simulations: $C_v$ Annual (1, 4, 5); Dry Year Fraction (1, 2, 5); Wetting Up Annual $C_v$ (1, 3); High Flow Discharge (2, 3); Skew Annual (2, 5); IQR ratio (3, 4); Peak Distribution Slope (1); Constancy (2); Slow Flow ($\alpha = 0.925$) (3); High Spell Duration $C_v$ (4); Falling Limb Density (4); and Flow Duration Curve (5) [where 1=GR4J; 2=SIMHYD; 3=IHACRES; 4=GR4JMOD; 5=SACRAMENTO].

- The techniques demonstrated in this chapter should be extended to other model structures where possible, particularly those that are regularly used in climate change impact assessments.

- Hydrologists should continue to explore methods for extracting information content out of available calibration data. Whereas most authors previously assumed that the poor performance of models under changing climate was
due to the model structures themselves (Section 5.2), this study suggests that improvements may be possible without changing the model structures. Even in cases where there is no problem with model structures, commonly used (sum-of-squares) calibration methods may fail (cf. Chapter 4).

- Further research should aim to examine the model-specific nature of signatures - i.e. why the replication of certain hydrologic signatures was associated with robust performance for some model structures but not others (Section 5.5.1).

- Future studies or projects aiming to produce simulations under changing climatic conditions should more carefully consider whether methods relying on random ensembles are appropriate (cf. Section 5.5.3).

5.6 Conclusions

The aim of Part 1 of this study was to search for ways to improve model calibration methods for changing climatic conditions, by analysing patterns relating metric values over a pre-change (wetter) calibration period and model performance over a post-change (drier) evaluation period. This aim has been fulfilled, with the identification of numerous hydrologic signatures and objective functions which could help to identify parameter sets that are more robust to changes in climate.

The hypothesis for the study as a whole was that there exists at least one metric with the property that optimising over a pre-change (wetter) period leads to improved model performance over a subsequent post-change (drier) period, compared to the reference calibration method. The results indicate that this hypothesis is likely to be true, and this is subject to confirmation in Chapter 6. A number of objective functions showed considerable promise for all models tested. For hydrologic signatures, multiple signatures were identified for which signature matching was associated with improved model performance as per this hypothesis. However, whereas each promising objective function tended to show favourable results across model structures, each promising hydrologic signature tended to show favourable results for some model structures and not others. Thus, signatures were more model
specific than objective functions. Further work into understanding these interactions was recommended.

The selected metrics imply multiple strategies for improving calibration methods for changing climates, including revising the common reliance on 'least squares' approaches in favour of 'absolute error' approaches (cf. Willmott and Matsuura, 2005; Chai and Drazler, 2014); and considering dynamics over a variety of timescales (not just daily) when calibrating models. These strategies have the potential to better prepare models for future climatic changes. The next step is guided search techniques such as optimisation (in the case of objective functions) or signature matching algorithms (in the case of hydrologic signatures) to confirm the initial findings of Chapter 5. This is explored in the next chapter.
Chapter 6

Improving calibration methods, 2: Comparison of Methods

6.1 Chapter Summary

It has been widely shown that rainfall-runoff models often provide poor and biased simulations after a change in climate, but evidence suggests existing models may be capable of better simulations if calibration strategies are improved. In this second part of a two-part study, we tested numerous calibration strategies, looking for methods that identify parameter sets that are robust to changes in climate, focusing on performance when evaluated over multi-year droughts. Various approaches were trialled, including hydrologic signature matching to over 200 different signature combinations (using the DREAM-ABC algorithm), optimisation of 8 different global objective functions (using the CMA-ES algorithm), and hybrid approaches blending global objective functions with signatures (using the Pareto approach AMALGAM). The results show significant improvements are possible compared to common ‘least squares’ approaches to calibration. Two notable methods were optimisation of the Refined Index of Agreement (based on sum-of-absolute-error, not sum-of-squared-error) and optimisation of a new objective function called the Split KGE (which

---

1This chapter has been submitted to the journal Water Resources Research and is currently under revision, as described in the preface.
weights each year in the calibration series equally). These objective functions obtained significantly better split sample results than common calibration methods like optimising the Nash Sutcliffe value. They also out-performed signature-based approaches, despite promising signature results from Chapter 5. We recommend future studies avoid ‘least squares’ approaches (such as optimising the NSE, RMSE or KGE) and adopt these alternative methods, wherever simulations of a drying climate are required.

The key points of this chapter are:

- ‘Least-squares’ approaches should not be used to calibrate models for a drying climate.
- Sum-of-absolute-error calibration approaches tend to select more robust parameter sets.
- Equally weighting all years in the calibration data tends to make calibration more robust.

### 6.2 Introduction

Rainfall-runoff models often provide poor simulations with high bias when applied in changing climatic conditions (Refsgaard, 1997; Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012, 2014; Singh et al., 2011; Saft et al., 2016a). This undermines their use for projecting impacts of possible future climate changes (as done by, eg. Bergström et al., 2001; Chiew and McMahon, 2002; Christensen et al., 2004; Fowler et al., 2007; Chiew et al., 2009; Vaze et al., 2011; Bosshard et al., 2013; Faramarzi et al., 2013; Hagemann et al., 2013; Forzieri et al., 2014). In response, researchers have recommended various courses, including placing limits on the degree of extrapolation required of rainfall-runoff models (cf. Vaze et al., 2010; Singh et al., 2011; Li et al., 2012) at least until model structures can be improved for changing climatic conditions (cf. Merz et al., 2011; Petheram et al., 2011; Coron et al., 2014). However, part of the problem may lie with calibration methods, as reviewed in Chapter 5. If
so, research to improve calibration methods could provide more robust simulations without the need to alter model structures.

By analysing patterns in large random ensembles of parameter sets, Chapter 5 of this study identified numerous promising metrics that could be used to improve calibration methods for changing climates. This chapter focuses primarily on calibrating rainfall-runoff models to these metrics using guided search techniques, that is, techniques not based on random ensembles. The focus of both parts is the transition from wetter conditions to drier conditions. The following paragraphs provide a review of calibration methods in general, with a focus on methods relevant to the present context. For a review of rainfall-runoff modelling under a changing climate, readers are referred to Chapter 5.

The earliest rainfall-runoff modelling studies used manual calibration. Early computers were not able to conduct large numbers of runs, so modellers would iteratively run the model, inspect the outputs, and fine-tune the parameter values manually. This approach had significant advantages, including the consideration of multiple (often competing) objectives simultaneously (Gupta et al., 1998; Boyle et al., 2000); the significant capacity of the human mind to recognise subtle patterns (Friedman and Tukey, 1974; Davtalab et al., 2017); the subjective consideration of possible data errors (Boyle et al., 2000); and the ability of a well-trained modeller to intuitively compensate for shortcomings of the model. Given its advantages, manual calibration continued to be used in some areas, notably in the preparation of flood forecasting models by the National Weather Service of the USA (Smith et al., 2003), but remains manually intensive (Boyle et al., 2000).

Although automatic approaches promised an objective and (theoretically) repeatable calibration method, numerous problems became clear. Early automatic methods required the modeller to choose a single mathematical function to quantify the ‘goodness’ of a model (a ‘global’ objective function), so that all the subtle information used to inform manual calibration was collapsed into a single number (Boyle et al., 2000; Gupta et al., 2008). Hydrologists commonly use least-squares based objective functions such as the Nash Sutcliffe Efficiency (NSE, |Nash and Sutcliffe, 1970;
Krause and Boyle, 2005), but the theoretical advantage of least-squares methods - namely, that they are consistent with Gaussian statistics (Sorooshian and Gupta, 1995) - is undermined by the fact that most studies do not check that the model residuals conform with Gaussian assumptions. In fact, errors are rarely homoscedastic due to the nature of flow measurements, among other reasons (Sorooshian and Dracup, 1980; Sorooshian and Gupta, 1995) and are rarely independent in time due to the inherent autocorrelation of model structural errors (Kuczera, 1983; Renard et al., 2011). In addition, model inputs such as rainfall tend to be subject to large errors (Andréassian et al., 2001). Methods to compensate for these tendencies are available (Sorooshian and Dracup, 1980; Kuczera, 1983; Kavetski et al., 2006a,b; Renard et al., 2011) but add extra dimensionality to the problem by requiring detailed error structures, with associated parameters, to be specified, and can be highly technical. These issues aside, the practical problem of finding the globally optimum region in the parameter space (e.g. given local optima) has prompted significant research into evolutionary based methods that can work in the highly non-linear objective space of rainfall-runoff models (e.g. Duan et al., 1992; Vrugt et al., 2008) and research to minimise the irregularities of this space by improving the design (Kavetski and Kuczera, 2007) and numerical implementation (Clark and Kavetski, 2010a) of rainfall-runoff models.

Given input and output uncertainties and imperfect methods of parameter set selection, the idea of choosing a single ‘best’ parameter set has been increasingly criticised (e.g. Beven and Binley, 1992; Kuczera and Mroczkowski, 1998) in favour of methods that consider an ensemble of different parameter sets. Numerous ensemble methods have arisen, including those based on random sampling of the parameter space followed by rejection of parameter sets based on pragmatic thresholds (Beven, 2006) and others based on a guided search of the parameter space that iteratively migrates through the parameter space in the direction of increasing performance Kuczera and Mroczkowski (1998); Vrugt et al., (2008). Both methods have the advantage that the ensemble of parameter sets can be used to estimate aspects of uncertainty in model predictions, although this is often debated (Beven and Binley,
6.2. INTRODUCTION

1992; Mantovan and Todini, 2006; Beven et al., 2008). In cases where a single parameter set is ultimately required, such as studies of very large samples of catchments or studies where catchment modelling is only part of a wider whole, the concept of data depth as defined by Tukey (1975) may be useful in selecting a parameter set from the ensemble. Data depth is a measure of whether an ensemble member is surrounded by other members in a multidimensional space; thus, parameter sets with higher depth are closer to the centre of the cloud of parameter sets, in parameter space. The identification of the ‘deepest’ set in an ensemble takes into account the position in parameter space of every other member in the ensemble, making it a better candidate to ‘represent’ the ensemble than the numerically optimum parameter set. Bárdossy and Singh (2008) demonstrated that parameter sets with higher depth were more transferable to other time periods.

Another difficult issue in model calibration is the choice of calibration function, given the issues with ‘least squares’ functions outlined above. Different functions weight data points differently, and the difficulties of choosing a single objective function have led hydrologists to trial multi-objective methods. These include methods which combine numerous objectives into a single objective (e.g. Zhang et al., 2008) and methods which can define the trade-off between two or more objectives (i.e. Pareto optimisation - Pareto et al., 1972; Yapo et al., 1998; Vrugt and Robinson, 2007; Efstratiadis and Koutsoyiannis, 2010). More recently, hydrologists have extended multiple-objective logic to include the concept of hydrologic ‘signatures’. A hydrologic signature is a statistic about the flow regime that can be calculated separately for the observed and simulated timeseries. The calibration goal is then to match simulated with observed values for one signature or a group of signatures. Guptha et al. (2008) argued strongly for the use of signatures because they are flexible, can be chosen to reflect aspects of the flow regime relevant to the study at hand, may be more suited to constraining highly parameterised models than single objective approaches (provided multiple signatures are used together), and have greater potential to diagnose which elements of the flow regime the model is failing to replicate. This last point, they argue, may also allow hydrologists to “determine those
components of the model ... [that] explain the discrepancy between computed and observed” (Gupta et al., 2008, p3806). Thus far, hydrologic signatures have proved useful in the context of the Prediction in Ungauged Basins (PUB) initiative as regionalised signature values can be informative during model calibration, sidestepping the difficulties of direct regionalisation of model parameters (cf. Bärdossy, 2007; Yadav et al., 2007; Balygina et al., 2009). Their utility in the context of traditional model calibration is developing (Westerberg et al., 2011; Hingray et al., 2010; Euser et al., 2013; Shafii and Tolson, 2015), with recent studies exploring Approximate Bayesian Computation (ABC) as a means of calibrating models to hydrologic signatures (Nott et al., 2012; Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014). The ABC method is discussed in further detail below in Section 6.3.3.

Returning to the context of changing climatic conditions, the development of calibration methods that maximise the robustness of parameter sets for projection under changing climates remains a clear need with much research still to be done, as reviewed in the previous chapter. Despite advances in some areas such as the selection of periods and systematic cross validation (eg. Coron et al., 2012), more research is required in many areas, such as the issue of choosing a calibration function. Most studies testing models over changing climates continue to use variants of least squares calibration functions, such as optimisation of NSE, despite the known tendency of such methods to underestimate variability of streamflow and sometimes mask high bias (Gupta et al., 2009). The IAHS workshop of 2013 entitled Testing simulation and forecasting models in non-stationary conditions defined a set of multiple evaluation measures relevant to changing climates, including metrics related to bias, low flows, and variability (Thirel et al., 2015b). However, the development of new calibration methods was not the focus of this workshop.

In summary, the following points are salient for this chapter: (i) there is a need for further research to develop calibration methods specifically for changing climatic conditions; (ii) the ability to consider multiple objectives simultaneously may be important, and hydrologic signatures may assist with this; (iii) the concept of ‘data depth’ may be useful in the context of changing climatic conditions and has
been shown to result in more transferrable parameter sets; and (iv) the theoretical advantage of adopting least-squares based objective functions (eg. NSE) is eroded by the tendency of modelling residuals to violate Gaussian assumptions.

This chapter builds on the lessons from the previous chapter. Specifically, numerous metrics were found (either objective functions or signatures) that appeared to be associated with better model performance under a change in climate (Section 5.5.5). However, the previous chapter based findings on random samples, not guided approaches, and such approaches were shown to have limitations (eg. Section 5.5.3). Based on the patterns observed in the previous chapter, this chapter tested different calibration methods with rainfall-runoff models applied in changing climates. The aim of this chapter is:

To identify a model calibration method, or methods, that provide improved rainfall-runoff model performance when applied under changing climatic conditions.

The hypothesis of both chapters is:

That there exists at least one metric with the property that optimising that metric over a pre-change (wetter) period leads to improved model performance over a subsequent post change (drier) period, compared to the reference calibration method.

Definitions of terms are the same as the previous chapter. As previously, performance refers primarily to the value of the Kling Gupta Efficiency (KGE - Gupta et al., 2009); however, in this chapter we supplement results with other evaluation metrics (Section 6.4.1 and Appendix D.7).

As noted in the previous chapter, another way of stating the hypothesis is that we believe that common calibration methods can be improved, in the context of changing climatic conditions. This chapter aims to rigorously test the potential improvements identified in Chapter 5.

### 6.3 Method

This study uses the same set of rainfall-runoff model structures (GR4J, SIMHYD, IHACRES, GR4JMOD and SACRAMENTO) and the 86 study catchments with
the same input data as Chapters 4 and 5. A large sample of catchments is important to maximise the chance of generality of results, and also to assess whether the results vary systematically with attributes such as climate, vegetation or physical characteristics like slope. The signatures and objective functions in this study were subsets of those used in the previous chapter, as summarised in Section 6.3.1. Subsequent sections present the calibration approaches adopted, including an explanation of why guided search techniques were used instead of random ensembles (6.3.2) and a description of the search algorithms adopted for each of the three calibration approaches (6.3.3). The final sections describe how results from the three methods were compared together on a ‘like-with-like’ basis (6.3.4); and how the methods were applied within the context of differential split sample testing (6.4.2).

6.3.1 Metric selection

The previous chapter developed shortlists of metrics regarded as ‘promising’ which are tested in this chapter. For completeness, the shortlists are repeated below. Metric definitions are in Tables 5.2 and 5.1 in the previous chapter.

**Objective functions:** Chapter 5 indicated the following six objective functions warrant further testing: Refined Index of Agreement; NSE\textit{square–root}; NSE\textit{fifth–root}; Split KGE, Split KGE 25 and Zhang Combination. In general, these objective functions provided promising results regardless of model structure. In addition to these objective functions, the NSE and KGE were also included since optimisation of these metrics is common in the literature, and thus they provide a useful point of reference. In particular, optimisation of KGE is used in this study as the ‘reference’ calibration method, used to represent common practice (as per Chapter 4).

**Hydrologic signatures:** Hydrologic signatures warranting further testing were different for each model structure. For each model structure, Chapter 5 identified four shortlisted signatures, listed below with numbers indicating which model structures each signature was associated with, where 1=GR4J; 2=SIMHYD; 3=IHACRES; 4=GR4JMOD; and 5=SACRAMENTO. The promising signatures were: C\textsubscript{v} Annual (1, 4, 5); Dry Year Fraction (1, 2, 5); Wetting Up Annual C\textsubscript{v} (1, 3); High Flow
6.3. METHOD

Discharge (2, 3); Skew Annual (2, 5); IQR ratio (3, 4); Peak Distribution Slope (1); Constancy (2); Slow Flow ($\alpha = 0.925$) (3); High Spell Duration $C_v$ (4); Falling Limb Density (4); and Flow Duration Curve (5).

6.3.2 Random ensembles or guided search techniques?

The method in Chapter 5 depended solely on large ensembles of randomly generated parameter sets, and analysis of patterns therein. In this chapter it would be possible to continue using randomly generated sets without recourse to guided search methods, as many researchers have demonstrated (e.g. Beven and Binley, 1992; Freer et al., 1996; Beven, 2006; Westerberg et al., 2011). However, this chapter focuses on guided search techniques because Section 5.5.3 in the previous chapter illustrated that a random ensemble may sparsely populate the area of interest within a multi-dimensional objective space, even when the ensemble is quite large. Based on these results, we believe methods that rely on random samples may occasionally fail to detect regions of the parameter space that produce robust simulations in changing climates, thus underestimating the ability of the model structure. It is acknowledged that guided techniques are also not guaranteed to find robust regions, as was discussed in Section 5.5.3.

6.3.3 Calibration approaches and algorithms

This chapter adopted calibration approaches as required for the type of metric in consideration. For the objective functions listed in Section 6.3.1, a single objective optimisation approach was used, as described below (Section 6.3.3). For the hydrologic signatures, a signature matching approach was used (Section 6.3.3). Section 6.3.3 describes a hybrid approach that utilised both types of metric (objective function and hydrologic signature). Each sub-section lists the context, algorithm used, and methodological details.
Single objective optimisation

Context: Single objective optimisation was used in this chapter to calibrate models to each of the eight objective functions listed in 6.3.1.

Algorithm: The adopted algorithm was the evolutionary optimisation algorithm CMA-ES (Hansen, 2006). This algorithm compares favourably with other common optimisers (Arsenault et al., 2014) and has previously been applied in hydrology (Peterson and Western, 2014, see also Chapter 4) and various other fields (Hansen, 2006). The output of CMA-ES is a single parameter set purported to correspond to the highest possible value of the objective function.

Method: For each of the eight objective functions separately, the CMA-ES optimiser was applied in each of the 86 study catchments, for each of the five model structures, giving a total of 3440 CMA-ES runs. Consistency of results was checked using the same method as in Chapter 4; namely, by running CMA-ES three separate times and if the optimum objective function value was not the same within 1%, the number of restarts (the only user-defined parameter in CMA-ES) was increased by one (starting from zero restarts). The process was then repeated until consistency was achieved.

Before proceeding to explain the other search methods, it is worthwhile to explain why other search methods were required at all. The reason is that CMA-ES (or any single objective optimiser) is not well suited to work with signatures. Signatures only characterise one aspect of the flow regime, such as the mean, the variability, or the recession index. It is entirely possible to exactly match one aspect and very poorly match others. Thus, as was shown in Figure 5.1b of the previous chapter, some parameter sets that match a signature have very poor scores in traditional metrics like the KGE. A key finding of the previous chapter was that some signatures had the property that robust performance over changing climate was correlated with the degree of signature matching, provided only the parameter sets with good KGE over the nondry (calibration) period were considered. To use this finding in practice, some method is required to screen out the sets with poor KGE over the nondry (calibration) period, and two such methods were trialled in this chapter. The first
(Section 6.3.3) was a hybrid signature-objective function approach which used an objective function (the KGE) to screen out the poorly performing parameter sets. The second approach (Section 6.3.3) used additional signatures to screen out the poorly performing parameter sets.

**Hybrid approach**

*Context:* The hybrid approach was applied to calibrate models to each of the shortlisted signatures listed in Section 6.3.1. However, it did not consider only the signature values; it considered both the match to a given signature (over the nondry (calibration) period) and the score in an objective function, the KGE (over the same period). The latter was used to filter out parameter sets that performed poorly over the nondry (calibration) period.

*Approach and algorithm:* A multi-objective Pareto approach (*Pareto et al.*, 1972) was adopted using the AMALGAM algorithm (*Vrugt and Robinson*, 2007). The Pareto Curve (*Pareto et al.*, 1972) defines the ‘trade-off’ between two objectives, in this case the error in signature matching (ideal score of zero) and the objective function score (ideal score of one). If parameter sets exist that match the signature value without sacrificing objective function score, such parameter sets will be on the Pareto Curve and can thus be identified using the Hybrid Approach. As explained by (*Vrugt and Robinson*, 2007), AMALGAM is an evolutionary scheme that combines strengths from other multiobjective schemes into a single robust algorithm and has been widely applied across a variety of fields (for hydrological applications, see *Wöbling et al.*, 2013, and Chapter 4 of this Thesis). As an evolutionary scheme, the user chooses the population size (here N = 100 as per Chapter 4); the output of AMALGAM is N parameter sets that together make up the Pareto Curve.

*Method:* For each of the five model structures, the AMALGAM algorithm was run separately for each of the four shortlisted (model-specific) hydrologic signatures (Section 6.3.1) for each of the 86 study catchments for a total of 1472 AMALGAM runs. In each case the objective function used was the reference metric of this study, the KGE. KGE was a natural choice given that the problem of parameter set screening
was originally framed in terms of KGE (cf. the previous chapter, Figure 5.1 and Section 5.5.2). Consistency of results was checked using the same method as in 4: namely, by running AMALGAM three times at 10,000 function evaluations, checking to see whether the curves were sufficiently close together in objective space using the rules outlined in Section 4.3.8, and re-running if required with increased function evaluations (up to a maximum of 80,000 function evaluations).

**Signature matching using Approximate Bayesian Computation**

*Context:* As with the hybrid approach, this method was applied to calibrate models to the shortlisted signatures listed in Section 6.3.1. However, differences with the hybrid approach included: (i) no reference was made to objective functions (ie. parameter sets were selected purely based on signature matching); (ii) more than one of the shortlisted signatures was sometimes used; and (iii) the shortlisted signature(s) were supplemented by extra signatures intended to screen out parameter sets that performed poorly over the non-dry (calibration) period. Details of signature selection are provided at the end of this sub-section.

*Approach and Algorithm:* Signature matching was undertaken using the framework of Approximate Bayesian Computation, or ABC (*Diggle and Gratton*, 1984). ABC is a method of model conditioning widely applied in the biological sciences (*eg. Pritchard et al.*, 1999) involving selection of parameter sets based purely on their match with one or more summary statistics (in this context, hydrologic signatures). As outlined by *Vrugt and Sadegh* (2013), ABC is a significant departure from the standard hydrological practice of step-by-step comparison of observed and simulated flows (as is done, for example, in calculating the NSE). Although non-guided ABC search algorithms exist (*eg.* the ABC-reject strategy described by Vrugt and Sadegh, 2013), we adopted the more efficient DREAM-ABC algorithm of *Sadegh and Vrugt* (2014) which is an evolutionary scheme implemented in the same package of code as the popular DREAM package (*Vrugt et al.*, 2008; *Vrugt*, 2016). As with standard DREAM, the output is not a single parameter set but an ensemble of ‘behavioural’ parameter sets. Behavioural sets all match the specified set of signatures to within
6.3. METHOD

a small tolerance $\varepsilon$. For this study $\varepsilon$ was defined in relative terms as ‘within 2%’ as described in Appendix D.2.

**Method:** The method for DREAM-ABC was more complicated and much of the detail is provided in the Appendices listed below. As mentioned previously, short-listed signatures were supplemented by additional signatures intended to screen out parameter sets that performed poorly over the nontidy (calibration) period. These additional signature combinations were sourced from an earlier study which identified the following signature combinations as commonly associated with high KGE values:

- Mean, $C_v$ daily and Catchment Lag A;

- Mean, $C_v$ daily and Slow Flow Index ($\alpha = 0.925$); and

- Mean, $C_v$ daily, Catchment Lag A and Slow Flow Index ($\alpha = 0.925$).

Note that this earlier study is heretofore unpublished and is therefore summarised in Appendix D.3.

It was not known a priori which of these three combinations was best and which of the four shortlisted signatures was best, so all possible combinations were tested, as follows. Considering the four shortlisted signatures for each model structure (Section 6.3.1), the list of all possible combinations of one, two, three or four of these signatures contains 15 possibilities. Pairing each of these with each one of the three combinations listed above gives 45 possible combinations per model structure. We also tested each of the three combinations with no shortlisted signatures added, for a total of 48 combinations. Running each of these combinations for each of 86 catchments and 5 model structures would require a total of over 20,000 DREAM-ABC runs. Due to the large number of function evaluations required by each DREAM-ABC run (Appendix D.2.1), this number of DREAM-ABC runs was not feasible. Therefore, we decided to still trial every possible combination, but to do so on a reduced set of 12 catchments. Rather than choose these catchments randomly, selection rules were used like those in Chapter 5, as outlined in Appendix D.4.
Appendix D.2 lists the DREAM-ABC settings adopted, based on recommendations of the original authors (Sadegh and Vrugt, 2014) where possible.

**Summary of calibration methods for comparison**

Summarising the above section, the three calibration methods examined in this chapter were:

a. Single objective optimisation using the CMA-ES algorithm. This was used to test all shortlisted objective functions. Every shortlisted objective function was tested for every model structure in every catchment;

b. Hybrid approach with signatures and KGE, using the AMALGAM algorithm. This method attempted to match one of the shortlisted signatures while at the same time attaining a high KGE in calibration. For each model structure, every shortlisted signature was tested in every catchment; and

c. Signature matching using the DREAM-ABC algorithm. This method required that the model simulations match the observed values of multiple signatures simultaneously. Large numbers of signature combinations were tested on a reduced set of catchments.

### 6.3.4 Comparability of methods

When comparing results from different search methods, difficulties arise in comparing ‘like with like’. For example, a great advantage of DREAM-ABC is that it produces an ensemble of parameter sets which can be used to estimate predictive uncertainty (Vrugt and Sadegh, 2013). In contrast, CMA-ES provides only one ‘optimum’ parameter set, so the outputs of these two methods are not directly comparable. Likewise, the output of AMALGAM - a Pareto Curve composed of multiple parameter sets - is not directly comparable to either of the other methods.

In this chapter, comparison was facilitated by choosing a single parameter set to represent a given DREAM-ABC ensemble or AMALGAM Pareto Front. Indeed, even if we focussed on only one method, this step would have been necessary to
facilitate comparison among large numbers of catchments and model structures. The methods of choosing ‘representative’ parameter sets were as follows:

- For AMALGAM outputs (consisting of a Pareto Front discretised as 100 parameter sets) the middle parameter set along the curve was chosen to represent the curve as a whole. This is demonstrated graphically in the results section (Figure 6.2). This was largely a pragmatic choice intended to avoid multiple arbitrary or subjective selection rules. Such rules would likely need to be defined individually for each signature tested, causing an unhelpful level of complexity.

- For DREAM-ABC outputs (consisting of an ensemble of behavioural parameter sets) the parameter set with greatest data depth, also known as half space depth (as defined by Tukey, 1975) was chosen to represent the ensemble as a whole. As explained in the introduction, data depth is a measure of whether an ensemble member is surrounded by other members in a multidimensional space. Identification of the ‘deepest’ set in an ensemble considers the position in parameter space of every other member in the ensemble, making it a better candidate to represent the ensemble than the numerically optimum parameter set. Furthermore, modelling results shown in the Appendix D.6 indicate that the deepest parameter set tended to systematically provide more robust simulations than the ensemble average (cf. Bárdossy and Singh, 2008), particularly for GR4J, GR4JMOD and SIMHYD, unless the model performance was generally poor in which case there was little difference. Data depth for each parameter set was quantified using the method of Rousseauw and Struyf (1998).

Finally, we note that in ABC, an additional class of outcome is possible: model structures can fail to match the specified signatures within the specified tolerance. This leads to a ‘null’ result where no behavioural ensemble is generated, so that no behavioural simulations are possible. The number of model failures are reported in the results section and the implications of this additional category of results are discussed in Section 6.5.3.
6.3.5 Differential split sample testing

All calibration methods tested in this chapter were evaluated using Differential Split Sample Testing. As outlined by (Klemes, 1986), Differential Split Sample Testing involves evaluating model performance over a period with conditions that are different to those in the calibration data. Here, this was defined by climatic conditions and achieved using the same period definition as in Chapters 4 and 5:

- The 7 driest consecutive years on record were used for model evaluation only and are referred to as the ‘dry (evaluation) period’. These years were chosen individually for each catchment based on streamflow.
- The remainder of the available timeseries was used as the calibration period and is referred to as the ‘nondry (calibration) period’. All cases of optimisation involved maximising objective function values calculated over the nondry (calibration) period. Likewise, all cases where signature values were matched refer to signatures calculated over the nondry (calibration) period.

6.4 Results

6.4.1 Results template

The main template for reporting results is repeated four times (Figures 1, 3, 5 and 6), so a summary of its main features is appropriate prior to reporting the results.

- The template presents boxplots of KGE values across all 86 study catchments (N=86), except for DREAM-ABC results (Figure 6.5) which for which N=12 (cf. Section 6.3.3);
- The top set of 10 boxplots refer to the calibration method optimisation of KGE, which is used as the reference method representing common practice. Shortlisted methods follow underneath.
- Given that the previous chapter found significant differences between model structures, the template shows results separately for each model structure,
indicated by boxplot colour.

- Given the focus on improving simulations over the dry (evaluation) period, the KGE values over the dry (evaluation) period, termed $KGE_{dry}$, are shown in bold boxplots. The KGE values over the nondry (calibration) period, termed $KGE_{nondry}$, are shown in faded boxplots.

- The x axis is linear in the range $-1 < KGE < 1$ and quasi-log for $-10 < KGE < -1$. Values less than -10 are shifted to -10 and the number of shifted instances is indicated in grey numbering.

Such plots are limited by the necessity to choose a single metric to quantify model performance (ie. the x-axis metric). In Figures 6.1 - 6.7, KGE is the chosen metric, but equivalent plots are provided in the Appendix, Figures D.7-D.11 using numerous other evaluation metrics, including the NSE, bias and several metrics recommended by Thirel et al. (2015a) for use in changing climates. Although the boxplots depict general patterns (eg. that one objective function performed better than another, on average across all catchments) a further limitation is that results in individual catchments are not shown (except as outliers), nor is the change in score for a given catchment when moving from one method to another. This is alleviated in Section 6.4.5 which contains scatter plots comparing results across three methods for individual catchments.

6.4.2 Results: objective functions

Figure 6.1 shows differential split sample test results for optimisation of the shortlisted objective functions. First consider results for the reference calibration method, optimisation of KGE. As expected, the calibration KGEs (faded boxplots) are uniformly high since the optimiser is aiming to maximise this score. However, the evaluation KGEs (bold boxplots) show some very low scores, indicating a significant drop in KGE when moving from the nondry (calibration) period to the dry (evaluation) period. Similar findings are apparent for the next method, optimisation of NSE. The calibration scores are not as high, which is expected since we have optimised to
a different (albeit related) metric to the one on the x-axis. However, the key result is
the KGE scores over the evaluation (dry) period (KGE\textsubscript{dry}), which are low, similar to
the reference method. Thus, optimising either the KGE or NSE provides relatively
low performance when evaluated over multi-year droughts. This is consistent with
the literature reviewed in Chapter 5 (eg. Refsgaard, 1997; Vaze et al., 2010; Merz
et al., 2011; Coron et al., 2012, 2014; Singh et al., 2011; Saft et al., 2016a).

Next let us consider the shortlisted objective functions. As expected, none of
the methods provide KGE scores as high over the nondry (calibration) period (faded)
as direct optimisation of the KGE. However, KGE scores in the dry (evaluation)
period are generally higher than the reference method for most shortlisted objective
functions. Among the best performing objective functions were the Refined Index
of Agreement, NSE\textsuperscript{square-\text{root}}, and the Split KGE, and these objective functions are
discussed in the following paragraphs.

The Refined Index of Agreement and NSE\textsuperscript{square-\text{root}} showed similar results in
Figure 6.1. A comparison of simulation bias (not shown here, see Appendix Figure
D.8) indicates that the Refined Index of Agreement provides less biased simulations
in evaluation than NSE\textsuperscript{square-\text{root}}, but the differences are relatively minor. These two
functions are also similar in formulation, as discussed in Section 6.5.1. For the Refined
Index of Agreement, the mean and median KGE\textsubscript{dry} across all 430 case studies was
0.519 and 0.635 respectively, up from 0.323 and 0.531 respectively for the reference
calibration method. Further, the number of instances where KGE\textsubscript{dry} was less than
zero was reduced from 20.2% (reference method) to 7.0% (optimising Refined Index
of Agreement). The figures in Appendix D.7 indicate that the Index of Agreement
and NSE\textsuperscript{square-\text{root}} performed well across a variety of evaluation metrics, providing
less biased simulations than the reference method for all model structures except
IHACRES, and better low flow replication for all model structures except GR4J.

It is instructive to compare the Split KGE results with standard KGE results.
Recall from the previous chapter that these are essentially the same metric, the only
difference being that the latter is calculated so that no year can have more influence
than any other year (cf. previous chapter Table 5.1). This relatively simple change
### 6.4. RESULTS

![Boxplots of differential split sample test results for single objective optimisation of various objective functions. All values plotted are KGE values (the reference metric). Boxplot colour denotes model structure. Faded colours indicate KGE scores in the nondry (calibration) period and bold colours denote KGE scores in the dry (evaluation) period. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +. Boxplots containing outliers <-10 are marked with a cross at -10 accompanied by a grey number indicating how many instances. This plot indicates significant potential for improvement over common calibration approaches.](image-url)

<table>
<thead>
<tr>
<th>Objective Function</th>
<th>Orange: GR4J (n=86)</th>
<th>Blue: SIMHYD (n=86)</th>
<th>Green: IHACRES (n=86)</th>
<th>Purple: GR4JMOD (n=86)</th>
<th>Red: SACRAMENTO (n=86)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind agree</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSE sqrt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSE 5th rt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SKGE 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.1: Boxplots of differential split sample test results for single objective optimisation of various objective functions. All values plotted are KGE values (the reference metric). Boxplot colour denotes model structure. Faded colours indicate KGE scores in the nondry (calibration) period and bold colours denote KGE scores in the dry (evaluation) period. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +. Boxplots containing outliers <-10 are marked with a cross at -10 accompanied by a grey number indicating how many instances. This plot indicates significant potential for improvement over common calibration approaches.
significantly improved the results. Across all models, the mean and median KGE_{dry} scores were 0.516 and 0.650, respectively, up from 0.323 and 0.531 respectively for the reference calibration method. The number of instances where KGE_{dry} was less than zero was reduced from 20.2% (reference method) to 8.4% (optimising Split KGE). The Split KGE 25 objective function also performed well, but hereafter we focus on the Split KGE instead of the Split KGE 25 for reasons discussed in Section 6.5.1. As with the Refined Index of Agreement, the Split KGE generally performed well against other evaluation metrics (Appendix D.7), providing simulations of relatively low bias and the closest replication of flow variability of any of the metrics tested.

In summary, these results strongly confirmed the shortlisted objective functions from the previous chapter. In general across the 86 study catchments, the shortlisted objective functions Refined Index of Agreement and Split KGE provided significantly improved simulations when evaluated over multi-year droughts.

6.4.3 Results: hybrid method

Before reporting hybrid method results across all case studies, two illustrative examples are presented (Figure 6.2) examining the trade-off between KGE and matching the signature C_v annual (cf. previous chapter Figure 5.1) for the GR4J model structure. In catchment 210011 (Figure 6.2a), the reference method (optimising KGE) produced a poor KGE score over the dry (evaluation) period, as shown by the colour of the ‘param. set adopted by reference method’ marker (KGE_{dry} = 0.28). However, this is improved by forcing a closer signature match while remaining on the Pareto curve, as shown by the gradient in colour moving leftwards along the curve starting from the right. Adopting the parameter set at the middle point on the curve (cf. Section 6.3.3) gives KGE_{dry} = 0.64, a higher value than the reference method. However, Figure 6.2b (catchment 407214) shows that the opposite can also occur. Note that the gradient in colour proceeds in the opposite direction to Figure 6.2a. This means that the KGE_{dry} value declines rather than increases in moving from right to left - the reference method gives KGE_{dry} = 0.84 but the middle of the curve gives 0.51. Thus, forcing the model to more closely match the C_v annual signature
Figure 6.2: Two examples of Hybrid Method results. The coloured curves are Pareto Curves indicating the trade-off between the KGE over the nondry (calibration) period and the match with the signature $C_v$ annual over the nondry (calibration) period. The colour refers to KGE in the dry (evaluation) period ($KGE_{dry}$) and is thus a measure of how robust the parameter set is (the most robust sets appear in yellow). Results are shown for (a) William River at Tillegra, New South Wales (Station 210011, catchment area 196.5km$^2$, mean annual rainfall 1190mm/year, runoff ratio 0.43); and (b) Creswick Creek at Clunes, Victoria (Station 407214, catchment area 306km$^2$, mean annual rainfall 686mm/year, runoff ratio 0.12). Each Pareto Curve is composed of one hundred parameter sets. The curve shading shows that a closer match with the signature leads to better KGE over the dry (evaluation) period in (a) but not (b).
Figure 6.3: Boxplots of differential split sample test results for the hybrid method. Formatting and interpretation of this plot is the same as for Figure 6.1 except for the following points. Because the shortlisted signatures vary depending on the model structure, the hybrid calibration methods are shown from best performing to worst performing, as judged by the median value in the boxplot. This ordering is determined separately for each model structure. This plot indicates that, in general, the hybrid did not lead to improved model performance when evaluated over multi-year droughts, compared to the reference calibration method (optimisation of KGE). Based on Pareto Fronts between KGE and matching with a single signature, as follows: (1) CvAnnual; (2) HighFlowDisch; (3) IQRatio; (4) CvAnnual; (5) SkewAnnual; (6) DryYearFrac; (7) SkewAnnual; (8) SlowFlow0925; (9) IQRatio; (10) DryYearFrac; (11) PeakDistSlope; (12) DryYearFrac; (13) WettingUpAnnCv; (14) HighSpellDurCv; (15) CvAnnual; (16) WettingUpAnnCv; (17) Constancy; (18) HighFlowDisch; (19) FallingLimbDens; (20) FDC.
6.4. RESULTS

odds with the strong patterns observed in Chapter 5, but closer inspection of results (as shown in Appendix D.5) suggests that the Figure 6.2b case is relatively common among the full set of 86 study catchments while being less common among the case studies used in Chapter 5 (cf. the previous chapter, Section 5.3.5). The hybrid method results suggest that closer signature matching may increase the robustness of simulations in some catchments for some model structures (such as those chosen for the previous chapter), but less so in general across all 86 catchments for a given model structure. The issue of representativeness of smaller samples of catchments is discussed further in Section 6.5.2. We now turn to the second of the signature methods (ABC) to determine if these findings hold for both methods trialled.

6.4.4 Results: Approximate Bayesian Computation

As with the hybrid method, we first present two example case studies (Figure 6.4). Both examples are plotted on the same axes as those used in Chapters 4 and 5 (note these are different axes to Figure 6.2), and the Pareto Curves (red) from Chapter 4 show the estimated limits of model performance in each case. For a selected DREAM-ABC run, each behavioural parameter set (ie. parameter set that matches all signatures in the set) is plotted as a dot, with colour indicating the data depth.

Figure 6.4a is an example of the DREAM-ABC method working favourably. The reference calibration method (left endpoint of Pareto curve) gives a KGE value over the dry (evaluation) period of 0.6 - while not poor, the Pareto Curve indicates room for improvement. In contrast, most of the DREAM-ABC behavioural sets are located in a favourable position close to the ‘maximum robustness’ position from the previous chapter. The exception is a thin tendril of parameter sets extending diagonally downwards; since these sets have relatively low data depth they are ignored when selecting a representative parameter set. The chosen set is quite close to the ‘robust region’ location from the previous chapter - a favourable result.

Figure 6.4b shows a less straightforward example, although still a positive outcome. The reference calibration method gives a poor KGE value over the dry
Figure 6.4: Two examples of DREAM-ABC results, for selected signature sets in (a) Happy Valley Creek at Rosewhite, Victoria (station 403214, catchment area 138km$^2$, mean annual rainfall 1193mm/year, runoff ratio 0.15); and (b) Bowna Creek at Yambla (station 401015, catchment area 274km$^2$, mean annual rainfall 670 mm/year, runoff ratio 0.05). DREAM-ABC behavioural sets are plotted as dots coloured by data depth. Pareto curves from Chapter 4 are marked as red diamonds. In both cases the DREAM-ABC deepest set outperforms the reference method in the evaluation (dry) period.

(evaluation) period (x axis value < 0.2), and so do many of the behavioural parameter sets. However, the behavioural ensemble covers a wide range of x-axis values, and some of the behavioural sets have KGE values greater than 0.6 for the dry (evaluation) period. In this case (cf. Appendix D.6), parameter sets with higher data depth tended to be located towards the right (higher KGE$_{dry}$), meaning that the representative (deepest) parameter set was relatively more robust than the ensemble as a whole. However, the ensemble itself was still relatively limited in its rightward shift, despite the inclusion of the shortlisted signature $C_y$ annual which was intended to encourage more robust results.

In contrast to these two favourable examples, the full set of results (Figure 6.5) showed that, in general, the DREAM-ABC method did not deliver more robust

---

2Signature combinations shown in Figure 6.5 were: (1) MeanDaily&$C_v$&CatchLagA&SlowFlow0925; (2) MeanDaily &$C_v$&SlowFlow0925; (3) MeanDaily&$C_v$&SlowFlow0925&SlowFlow0925&IQRratio; (4) MeanDaily&$C_v$&CatchLagA&HighSpellDurC&IQRratio; (5) MeanDaily&$C_v$&CatchLagA&DryYearFrac; (6) MeanDaily&$C_v$&SlowFlow0925; (7) MeanDaily&$C_v$&CatchLagA; (8) MeanDaily&$C_v$&CatchLagA&SlowFlow0925&IQRratio; (9) MeanDaily&$C_v$&CatchLagA&Annual; (10) MeanDaily&$C_v$&CatchLagA; (11) MeanDaily&$C_v$&CatchLagA; (12) MeanDaily&$C_v$&SlowFlow0925&Constancy; (13)
simulations than the reference method. Figure 6.5 is presented similarly to Figure 6.3, with higher performing methods located towards the top. Of the 48 signature combinations tested, only the top five are shown in Figure 6.3. Note that the sample sizes in Figure 6.5 were much smaller (N=12 rather than N=86), being affected by the selection rules outlined in Appendix D.4 (cf. Section 6.3.3). The left-hand column of Figure 5 contains the number of catchments (out of 12) in which the model structure failed to match the signatures in the combination. Since such failures meant a behavioural ensemble did not exist, it was not possible to produce simulations; therefore, affected boxplots are based on reduced samples of less than 12 catchments.

DREAM-ABC performed best for the IHACRES model structure, with significant improvement across the twelve example catchments and the removal of some strongly negative KGE_{dry} values. However, for the other model structures there was little improvement compared to the reference method.

In terms of model failures, the IHACRES model proved most able to match multiple signatures, with a low failure rate for most combinations (zero, one or two failures out of 12). In contrast, GR4J proved least able to match multiple signatures, with failure rates of up to 6 catchments out of 12. This is likely related to model parsimony and the difficulty of matching up to 8 signatures simultaneously with only 4 degrees of freedom (free parameters).

To finish the results section, Section 6.4.5 below compares results across all methods tested. Given that the other methods were run for all 86 catchments, extra DREAM-ABC runs were conducted to extend results to all 86 catchments. This was done only for the two best performing combinations of each model structure from Figure 6.5. Thus, in Section 6.4.5 the plots for DREAM-ABC are based on 86 rather than 12 catchments.

\begin{footnotesize}
\begin{tabular}{ll}
MeanDaily&C_a,C_p,CatchLagA,SlowFlow\textsubscript{0925}&IQRratio; & MeanDaily&C_a,CatchLagA,SlowFlow\textsubscript{0925}; \\
\end{tabular}
\end{footnotesize}
CHAPTER 6. IMPROVING CALIBRATION: METHOD COMPARISON

Figure 6.5: Boxplots of differential split sample test results for the DREAM-ABC method. Formatting and interpretation of this plot is the same as for Figure 6.3 except for the following points. The sample size is reduced compared to previous figures as per Section 6.3.3. Model failure statistics are included as crosses in the left hand column. For each model structure, the display is ordered so that better performing signature combinations appear in the boxplots towards the top. Only the best 5 signature combinations are shown (out of 48) for each model structure. For cases where boxplots appear out of order, note the following: ordering is by the 6th highest value, which is approximately the median of the boxplot only in the case of zero model failures, since model failures are counted as \(-\infty\) in the ordering but are not included in the boxplot. This plot indicates that, in general, the DREAM-ABC method did not lead to improved model performance when evaluated over multi-year droughts, compared to the reference calibration method (optimisation of KGE). Refer to the footnotes on the preceding pages for the signature combinations shown.
6.4. RESULTS

6.4.5 Results: comparison of methods

Figure 6.6 compiles the best results for each of the classes of method, into a single plot to facilitate comparison. The two optimisation methods shown (optimisation of the Refined Index of Agreement; and optimisation of the Split KGE) clearly provide superior results compared to the any of the other methods, as judged by KGE values over the dry (evaluation) period.

Figure 6.7 plots values of KGE in the evaluation (dry) period for the reference method against the two best calibration methods identified above: optimisation of the Refined Index of Agreement (Figure 6.7a) and optimisation of the Split KGE (Figure 6.7b). While these two methods did not always improve upon the reference method, the magnitude of change among cases of decline was usually quite modest relative to the magnitude of change among cases of improvement. Across all five model structures:

- For the Refined Index of Agreement, the $\text{KGE}_{\text{dry}}$ value increased (compared to the reference method) in 269 case studies and decreased in 161 case studies. Among cases of increase, the average change in KGE was 0.41 units; among cases of decrease, the average change in KGE was 0.16 units.

- For the Split KGE, the $\text{KGE}_{\text{dry}}$ value increased (compared to the reference method) in 302 case studies and decreased in 128 case studies. Among cases of increase, the average change in KGE was 0.33 units; among cases of decrease, the average change in KGE was 0.15 units.

From Figure 6.7, the improvements are greatest in cases where the reference method does relatively poorly ($x$ axis $< 0.5$). For cases where the reference method does relatively well ($x$ axis $> 0.5$) there is little difference between the mean result across catchments for reference and comparison methods. The Appendix indicates that the improved KGE scores in evaluation are primarily due to less bias in simulations (Figure D.8) and, in the case of Split KGE, less overestimation of variability in flows (Figure D.10).
CHAPTER 6. IMPROVING CALIBRATION: METHOD COMPARISON

Figure 6.6: Boxplots of differential split sample test results showing comparisons among different method types. For DREAM-ABC (bottom), note the inclusion of model failure statistics (crosses in left hand column). All values plotted are KGE values (the reference metric). Boxplot colour denotes model structure. Faded colours indicate KGE scores in the non-dry (calibration) period and bold colours denote KGE scores in the dry (evaluation) period. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +. Boxplots containing outliers < -10 are marked with a cross at -10 accompanied by a grey number indicating how many instances. This plot indicates that, in general, optimisation of the Index of Agreement or Split KGE each provide superior results compared to the hybrid and DREAM-ABC approaches, and compared to the common methods of optimisation of KGE and NSE. Hybrid methods used (1) $C_v$ Annual; (2) HighFlowDisch; (3) IQRratio; (4) $C_v$ Annual; (5) SkewAnnual. DREAM-ABC runs used (6) MeanDaily&$C_v$&SlowFlow0925; (7) MeanDaily&$C_v$&CatchLagA; (8) MeanDaily&$C_v$&SlowFlow0925&IQRratio; (9) MeanDaily&$C_v$&CatchLagA&C_vAnnual; (10) MeanDaily&$C_v$&CatchLagA.
Figure 6.7: Comparison between methods on a catchment-by-catchment basis (N=86 for each model structure). In each case the reference method (x axis) is compared with a shortlisted method (y axis). The plotted values are KGEs in evaluation from differential split sample tests. KGE values < -10 are shifted to -10 and marked with a circle.

Figure 6.7 also shows the spread of results for various catchment characteristics (smaller plots). For this set of catchments, high rainfall (long term average > 1500 mm/year) locations show negligible benefit from the alternative objective functions; using KGE as objective function generally leads to good results in these catchments (sub-figures a-ii and b-ii). In contrast, catchments less than 1500 mm/year show significantly improved results from using Index of Agreement or Split KGE compared to KGE. For this catchment sample, using Index of Agreement or Split KGE as objective function provides improved split sample results regardless of how steep the catchment is (sub-figures a-iii and b-iii) and regardless of how forested (sub-
figures a-iv and b-iv). We also compare results based on how severe the drought was (Figures a-v and b-v), as measured by flow reductions in the dry (evaluation) period relative to the nondry (calibration) period. In general, swapping the objective function from KGE to either Index of Agreement or Split KGE provides improved split sample results regardless of drought severity, but benefits are relatively greater in locations where the drought was more severe.

In summary, the Refined Index of Agreement and the Split KGE were the two objective functions that provided the best performance. They performed significantly better in split-sample testing than the NSE, KGE or NSE with bias penalty, regardless of model structure, catchment slope, or vegetation. For all except the highest rainfall locations tested (ie. mean annual rainfall > 1500mm/year), results indicated significant benefits in adopting these objective functions over least-squares type functions.

6.5 Discussion

6.5.1 Discussion of successful calibration strategies

The results presented in the previous section provide a strong case for improvement over commonly used calibration methods such as optimisation of NSE or optimisation of KGE. In this section we discuss the reasons underlying the success of various calibration strategies: optimisation of the Refined Index of Agreement; NSE\texttextsuperscript{square-root}; Split KGE; and Split KGE 25. As discussed in the introduction, through the squaring of errors on each timestep, least squares methods cause the influence of large errors to become larger, which tends to emphasise high flow days because model error and measurement error usually increase for higher values of flow (heteroscedasticity; cf. Legates and McCabe, 1999; Krause and Boyle, 2005; Criss and Winston, 2008). Figure 6.4 shows graphically the consequences of this tendency. In the calculation of sum of squares over the year 1975 (Figure 6.4b), only the two highest flow peaks influence the calculations, making the cumulative timeseries appear as a sharp step function, and causing the blue parameter set to be considered superior to the red
6.5. DISCUSSION

parameter set. Sum of absolute errors (as in the Index of Agreement) still considers the high flow days (Figure 6.4c), but the step function is more rounded and the inter-peak periods have non-zero gradient, indicating that the mid-low flows on either side of the peaks are also considered. This means that the red parameter set is considered superior, acknowledging its closer tracking of recessions and subsidiary peaks. We suggest that this more balanced consideration of the hydrograph is important for preparing a model for drier climatic conditions, as information on physical processes relevant to catchment drying may be contained in mid- and low- flow periods, in addition to high flow periods.

The NSE_{square-root} and the Refined Index of Agreement are quite similar, leading to similar (but not identical) results in Figure 6.1. Taking the square root of flows tends to counteract the effect of squaring the errors, so that the points of the previous paragraph broadly apply to the NSE_{square-root} also. Prior studies have used the square root transform to stabilise the variance prior to application of least squares measures (eg. Engeland et al., 2005), and the results of this study suggest that this practice does select more robust parameter sets. Note that NSE_{square-root} and the Refined Index of Agreement are not identical; differences arise because |a - b| ≠ (\sqrt{a} - \sqrt{b})^2 for most a and b. In terms of results, the main difference between NSE_{square-root} and Refined Index of Agreement is that the latter produced less biased simulations in evaluation (Appendix D.8).

Next, consider the Split KGE, which (as per Table 5.1) is calculated by splitting the timeseries into individual years, calculating the KGE value for each year in isolation, and then taking the average of these annual values. Like in Figure D.1, the effect is to ‘even out’ the influence of parts of the calibration timeseries, but the process happens on a timescale of years (rather than days for the Refined Index of Agreement). A given wet year may have a very high influence on the standard KGE score, but with the Split KGE the influence is limited to 1/N (where N is the number of years in the calibration period). Conversely, dry years may be ignored by the standard KGE - despite potentially containing the most relevant information for a drying climate - but with the Split KGE their influence is guaranteed to be
Figure 6.8: (a) Observed and GR4J simulated flow during part of a calibration pe-
riod, for two parameter sets: parameterisation by optimising KGE_{nondry} (parameter
set 1) and parameterisation by optimising Index of Agreement_{nondry} (parameter
set 2). The example is Currambene Creek at Falls Creek (216004, catchment area
93.5 km², mean annual rainfall 1130mm/year, runoff ratio 0.20). (b) cumulative evo-
lution of the sum of squared errors over the same period. (c) cumulative evolution
of the sum of absolute errors over the same period. Sum of squared errors is very
sensitive on days of high flow, but effectively ignores other days, as indicated by
the flat gradient in between floods. The sum of absolute errors is still sensitive to
floods but also considers lower flows.

1/N. Figure D.2 shows a practical example. The observed flows in each year are
provided (top) to indicate wet and dry years. The simulated flows are not directly
plotted, but the daily KGE value for each year in isolation is given, for parameter set
1 (chosen when the standard KGE is the objective function) and for parameter set
3 (chosen when the Split KGE is the objective function). The colour coding shows
that using the standard KGE as the objective function provides good performance
during most wet years but largely ignores poor performance during the dry years,
Figure 6.9: Annual observed flow (top) with tabulated KGE values for 41 separate calendar years, for two parameter sets: parameter set 1 is the same as Figure D.1 (optimising $KGE_{nondry}$); parameter set 3 is from parameterisation by optimising Split $KGE_{nondry}$. The case study is the same as in Figure D.1 (GR4J in 216004). Optimisation of standard KGE gives superior values during wet years while largely ignoring poor performance in the drier years of the calibration period, whereas the Split KGE considers all years equally. In terms of conventional KGE values, parameter set 1 attained $KGE_{nondry}$ (calibration) period = 0.85 and $KGE_{dry}$ (evaluation) period = -0.06, and parameter set 3 attained $KGE_{nondry}$ (calibration) period = 0.59 and $KGE_{dry}$ (evaluation) period = 0.61.

particularly towards the end of the calibration period. The Split KGE, in contrast, must consider each year as equally weighted, leading to more accurate simulations of most dry years in the calibration period, and subsequently better performance in the dry (evaluation) period.

Finally, we discuss the Split KGE 25 objective function, which is similar to the Split KGE except that it is not an average of all years, but is the average of the lowest 25% of annual KGE values. The Split KGE 25 obtained favourable results of similar standard to the Split KGE (Figure 6.3.3). Studies that adopt the Split KGE 25 should consider the possibility that valuable information in the calibration period may be ignored, due to focus on the years of lowest model performance. If observational error is greater in particular years for some reason, the Split KGE 25 may become unduly focussed on matching such periods even if more informative data are available in other years. This is particularly the case where calibration periods
are relatively short (eg. Coron et al., 2012). Therefore, in general, we recommend
the Split KGE in preference to the Split KGE 25.

6.5.2 Issues arising from judgements based on sub-samples of catch-
ments

Previous authors have argued the benefits of broad scale understanding based on
large samples of catchments (eg. Gupta et al., 2014). Accordingly, in this chapter
we adopted a relatively large sample (86 catchments). In the previous chapter, sub-
samples were used, and a comparison of the two parts provides interesting lessons
on the risks of focussing on limited sets of catchments.

In the previous chapter the set of 86 catchments was reduced to between 18
and 28 depending on model structure. As explained in Section 5.3.4, it was logical to
focus on catchments where models calibrated using the reference method performed
well during wetter periods but poorly during drier periods, and to ignore case stud-
ies where the Pareto analysis of Chapter 4 indicated that no robust parameter sets
existed within the model structure. This method was valuable, making it easier to
discern patterns from noise by focussing on catchments where patterns were most
likely to exist. However, the subsequent results in this chapter confirmed some Chap-
ter 5 findings (eg. regarding objective functions), while proving others to be false,
or at least, not generally true (eg. regarding hydrologic signatures). For example,
Figure D.4 shows, for the GR4J model structure and the C6 annual hydrologic sig-
nature, that the selection criteria used for Chapter 5 tended to preferentially select
catchments where the hybrid method worked, leading to the false expectation that
the method would work across the full sample of 86 catchments. Thus, the results
in this chapter emphasise the importance of testing on a large set of catchments.

Although the set of 86 catchments was relatively large, we note that they were
still only from one continent (Australia), and temperate climate types (Figure 2.1).
Thus, further research is recommended to determine whether the general conclusions
reached in this chapter hold for other continents and climate types.
6.5.3 Consideration of ABC as a calibration tool

Based on the results shown, we do not recommend the use of ABC to calibrate rainfall-runoff models for drying climates. However, the potential uses of ABC in hydrology extend beyond the focus of this chapter, so it is appropriate to provide a brief discussion drawing on the experience of applying ABC, and specifically the DREAM-ABC algorithm.

As noted, ABC introduces an additional category of result: model failure. Models fail if no parameter set is found that can match the signatures to the required tolerance (Section 6.3.3). Far from being a drawback, this could be seen as a positive feature. Numerous authors (e.g. Seibert, 2001; Schaefl and Gupta, 2007; Ritter and Muñoz-Carpena, 2013) have discussed the lack of objective standards in hydrological modelling due partly to the common use of sliding scales in model evaluation (e.g. NSE values) that provide no inherent threshold of acceptability. Vogel and Sankarasubramanian (2003) suggested that a useful ‘first test’ of a model structure (i.e. pre calibration) might be its ability to match system-relevant signatures; DREAM-ABC applied in this way may be useful for reducing a large list of model structures to a shortlist (cf. Euser et al., 2013), and/or confirming the suitability of a model structure for a given context prior to a dedicated calibration process (Martinez and Gupta, 2010). We also note the suggestions by Gupta et al. (2008) that hydrologic signatures be used to support ‘diagnostic’ model evaluation that can objectively determine the aspect(s) of the flow regime that a model is failing to match. While this idea holds promise (e.g. Hrachowitz et al., 2014), we note (cf. Appendix D.1) that, guided by DREAM-ABC, every model structure tested in this chapter could match any one of the 36 hydrologic signatures, provided it wasn’t required to simultaneously match any other signatures. Thus, a model structure’s failure to match a signature must be put in the context of what other signatures the model is also required to match, or what other calibration objectives are included.

DREAM-ABC has numerous other benefits relative to optimisation approaches. A significant benefit is its ability to facilitate uncertainty analysis. Standard optimi-
sation cannot be used for uncertainty analysis because it gives only one parameter set. Many methods that do provide an ensemble of parameter sets can only be defensibly used for uncertainty analysis if model residuals are consistent with the assumptions inherent in the likelihood function (Sorooshian and Gupta, 1995), and complicated error models are sometimes required to achieve this (eg. Kavetski et al., 2006a,b; Renard et al., 2011). In contrast, and as argued by Vrugt and Sadegh (2013), ABC provides an ensemble suitable for uncertainty analysis while (arguably) retaining conceptual simplicity. A further benefit is that DREAM-ABC can work directly with regionalised values of hydrologic signatures (cf. Yadav et al., 2007), although the version of DREAM-ABC used here does consider uncertainty in hydrologic signature values, which may be more of a limitation in regionalisation studies than in the present context.

ABC could be criticised because the choice of signatures is subjective, which effects the choice of parameter sets. Again, this could be rephrased as a benefit since the modeller can choose the signatures that are most relevant to the problem at hand (Vogel and Sankarasubramanian, 2003), not unlike the development of context-specific objective functions (eg. Pushpalatha et al., 2012).

Future users of DREAM-ABC may benefit from the adopted algorithm settings listed in D.2.1 that were guided by sensitivity testing as outlined in Appendix D.2.2.

6.5.4 Research Questions arising from this chapter

Numerous potential research questions arise from this work. As already mentioned:

- Future work could examine whether the general conclusions reached in this chapter hold for continents other than Australia and climate types other than humid;

- Future work could combine the ideas expressed in the Refined Index of Agreement and the Split KGE into a single objective function; and

Three further research areas are recommended. Firstly, this chapter focussed on calibration of models for drying climatic conditions, but some areas of the world
may see increased rainfall in future (Covey et al., 2003). Thus, further research is required to confirm if the calibration methods identified here provide robust simulations in the case of wetting, and if not, similar analyses to the previous chapter might identify calibration methods that do.

Secondly, this chapter focussed on five rainfall-runoff model structures commonly used in Australia and abroad. The two recommended objective functions provided consistently better results than the reference calibration method across all five of these model structures, providing some confidence that the objective functions may work well across other, untested, structures. Nonetheless, it is recommended that future studies applying these calibration methods with other model structures conduct their own split sample testing to confirm the suitability of these calibration methods, where possible.

Lastly, the utility of data depth has generally been confirmed by this study (cf. Appendix D.6) and further data depth research in hydrology is recommended (cf. Bárdossy and Singh, 2008). This study demonstrated that the deepest parameter sets in DREAM-ABC ensembles tend to provide more robust simulations in changing climate than the ensemble as a whole, although this advantage tends to be model specific (cf. Appendix D.6). Future research may apply data depth to ensembles generated in other ways. For example, a possible improvement upon the use of the Refined Index of Agreement or the Split KGE as demonstrated here with CMA-ES, might be to generate ensembles of parameter sets with high values in these metrics and then adopt the deepest parameter sets in these ensembles, rather than the numerically optimum parameter set.

6.6 Conclusions

As described in the introduction, this chapter aimed to identify a model calibration method, or methods, that provide improved rainfall-runoff model performance when applied under changing climatic conditions. This aim was fulfilled, and two different classes of method were particularly effective.
**CHAPTER 6. IMPROVING CALIBRATION: METHOD COMPARISON**

- Sum-of-absolute-error methods such as the Refined Index of Agreement of Willmott *et al.* (2012). For example, the average KGE score in evaluation across 86 catchments during differential split sample testing was 0.519 when models were calibrated by optimising the Refined Index of Agreement, compared with 0.323 when optimising the KGE directly. The practice of calculating the NSE on the square root of flows (termed NSE\textsubscript{square-root} in this chapter) is similar to a sum-of-absolute-error approach, and NSE\textsubscript{square-root} provided similar evaluation results to the Refined Index of Agreement but with slightly more bias.

- Methods which weight each year in the calibration series equally to limit the influence of wet years in calibration data and ensure dry years are not ignored. For example, the median KGE score in evaluation during differential split sample testing was 0.516 when models were calibrated by optimising Split KGE, compared with 0.323 when optimising the KGE directly.

These results suggest that there is information in calibration data that is not fully exploited by common calibration methods, and confirms the hypothesis that there exists at least one metric with the property that optimising over a wetter period leads to improved model performance over a subsequent drier period, compared to the reference calibration method (optimisation of KGE).

The success of the methods listed above was consistent across the five rainfall-runoff models tested. The study was limited to 86 humid catchments in southern Australia, and the testing was limited to the case of climate drying rather than wetting. Thus, further testing is required to confirm the generality of results in other locations, across model structures and for other types of climatic change.

We recommend future studies avoid ‘least squares’ approaches (such as optimising the NSE, RMSE or KGE) and adopt these alternative methods, wherever simulations of a drying climate are required. Continued research is needed on the extraction of information content from calibration data, alongside efforts to improve models. Whereas some studies previously assumed that the poor performance of models under changing climate was due to the model structures themselves, this
6.6. CONCLUSIONS

chapter demonstrated that improvements are possible without changing the model structures. Continued research may uncover further avenues for improvement. This chapter should encourage future modellers to employ a range of methods to extract information from data, rather than relying on commonly used methods of calibration.
Chapter 7

Framework for improving rainfall-runoff modelling under changing climate

7.1 Chapter Summary

Rainfall-runoff models often perform poorly when climatic conditions change. Recent literature has suggested tools to evaluate models under such change, usually using variants of the Differential Split Sample Test (DSST). There is less guidance, however, on what to do when a model fails these tests. Does failure in a DSST imply that changes are needed to a model structure, or to how models are initially calibrated, or both? If changes are made, how should they be evaluated? This chapter provides guidance for answering these questions within a framework based on Pareto optimality, where modelling objectives are set over multiple historic periods with contrasting climatic conditions. This approach allows cases of DSST failure to be categorised as either: (a) cases where no parameter set can meet modelling objectives in all periods, indicating the need for structural changes, improved data or time-varying parameters; or (b) cases where modelling objectives are attainable by

\footnote{This chapter has been submitted to the journal Water Resources Research and is currently under review, as described in the preface.}
the model structure, but the DSST calibration method failed to find the right parameter set(s). We outline steps to follow for each of the above categories, some of which can be populated by existing techniques. For some steps, we design new techniques such as the analysis of ‘drift’ in hydrologic signature error with climatic conditions to diagnose structural problems. We demonstrate the framework in an Australian catchment using the IHACRES model structure, and discuss the limitations of inferring future hydrologic processes from historic data. This framework aims to guide model improvements and foster greater confidence in hydrological projections.

The key points of this chapter are:

- A model improvement framework is presented, applicable to models failing split sample testing.
- Tests based on Pareto optimality are used to decide if model structural changes are required.
- Climatic ‘drift’ in hydrologic signature error is analysed to diagnose structural problems.

7.2 Introduction

Understanding how river catchments respond to changes in climatic forcing is important for projecting future water availability for human and environmental needs. Future climatic changes may go beyond the variability of the past (Covey et al., 2003; Forster et al., 2007; Meehl et al., 2007; Milly et al., 2008) and be amplified by hydrologic systems (eg. Vogel et al., 1999; van Dijk et al., 2013; Saft et al., 2015). Rainfall-runoff models are key tools for quantifying these non-linear responses, and are commonly used to estimate streamflow resulting from climate model projections (eg. Chiew et al., 1995, 2009; Bergström et al., 2001; Christensen et al., 2004; Faramarzi et al., 2013; Forzieri et al., 2014). Given their capacity to inform our understanding of the implications of emissions scenarios and human adaptation on water resources and ecosystem services, it is critical to ensure rainfall-runoff models
7.2. INTRODUCTION

can provide robust simulations under change, in line with the wider current IAHS emphasis on ‘change in hydrology and society’ (Montanari et al., 2013).

A typical assumption of climate change impact studies is that rainfall-runoff models calibrated on historical data are valid to simulate the future, even if at times future conditions are unlike those in the past, which is increasingly criticized as inappropriate (Clarke, 2007; Wagener et al., 2010). In cases where future variability goes beyond past experience, it is not possible to directly test the validity of this assumption, but a possible test is to split up the historic data based on climatic conditions, and calibrate models to a subset of historic climatic variability (Klemeš, 1986; Refsgaard, 1997; Vaze et al., 2010). Model performance over a different set of climatic conditions then provides an independent test of whether the model can operate under change, an example of the Differential Split Sample Test (DSST, cf. Refsgaard et al., 2014).

A key framework for tests of this type was proposed by Klemeš (1986), who included the DSST in a wider scheme of model evaluation strategies suitable for models applied in contexts such as changes in climate, changes in landuse, and changes in location (ie. the Prediction in Ungauged Basins problem - Blöschl et al., 2013). Models that passed such tests were referred to as ‘operationally adequate’ by Klemeš (1986, p13), who characterised the aim of the testing scheme as ‘merely to assess the performance of the model in situations as close as possible to those in which [it is] supposed to be used in practice’ and noted that passing such tests is a ‘necessary, rather than a sufficient, condition for model adequacy’ (Klemeš, 1986, p13). In this chapter, ‘DSST’ hereafter refers to Test 2a from Klemeš (1986), based on temporal changes in climatic conditions, not landuse.

In practice, rainfall-runoff models often perform poorly in the DSST, exhibiting substantial reductions in performance when applied in climatic conditions that are different to the calibration period. For example, Refsgaard and Knudsen (1996) applied three model structures of various complexity to catchments in Zimbabwe, noting that the model performance was poor in the driest years in the record, even for more physically based models. Vaze et al. (2010) tested four rainfall-runoff model
structures on 61 catchments in Australia, noting that model performance tended to
decline in proportion to the difference between periods in climatic variables (such as
rainfall), and that the decline was greater if the direction of change was from wetter
to drier conditions. Similar results have been confirmed by many other studies (eg.
Coron et al., 2012, 2014; Seiller et al., 2012; Seiller and Ancill, 2015; Broderick et al.,
2016; Saft et al., 2016a).

These DSST failures have prompted efforts in various areas, including advances
in techniques of model calibration/evaluation; research into changes in physical pro-
cesses with changing climatic conditions; and advances in methods for diagnosing
model structural deficiencies. With regards to the first category, Coron et al. (2012)
suggested the Generalised Split Sample Test, which conducts numerous DSSTs on
the same set of data via a sliding window in time. Each possible period of a chosen
length (they adopted 10 years) is used as a calibration period in turn, thus removing
the subjectivity of choosing the calibration period and allowing more insight into
the relation between model performance and climatic variables. In a similar vein,
Thivel et al. (2015b) argued that multiple time periods and multiple complementary
performance metrics should be evaluated in each test, including key statistics such
as the mean and variability of runoff values.

Second, the DSST failures have raised questions about possible changes in
physical processes caused by changing climatic conditions. Should such changes be
considered ‘within scope’ of conceptual rainfall-runoff models? For example, recent
advances in the characterisation of root zone storage suggest that this storage co-
evolves with vegetation to bridge droughts with return periods of 10-40 years (Gao
et al., 2014; cf. Troch et al., 2015) and might rapidly change following deforestation
(Nijzink et al., 2016). The key question is whether root zone changes could also occur
in response to changing climatic conditions, either through adaptation of individuals
or selective pressures among individuals and/or species to favour plants with different
rooting characteristics. This is one example of a possible change in relevant physical
processes. If such changes are occurring and are not represented by rainfall-runoff
models, then it is logical that the performance of such models will decline when these
7.2. INTRODUCTION

changes occur, and calibrated values of some parameters may drift systematically
with changing climatic conditions to compensate (cf. Wilby, 2005; de Vos et al.,
2010; Merz et al., 2011; Brigode et al., 2013). The modelling framework presented
in this chapter assumes that all such environmental changes should be regarded as
within-scope, if a model is to be used in climate change impact studies. This is not
limited to vegetation dynamics, but could include a variety of other processes, eg.
runoff implications of climate-induced groundwater decline (Hughes et al., 2012).

Regarding the third category, methods for diagnosing structural deficiencies
may provide valuable insight into potential model changes. Data-based model
improvement is central to hydrological simulation and has been demonstrated for many
models and locations (eg. Ambrose et al., 1996; de Vos et al., 2010; Westra et al.,
2014). However, it has rarely been done systematically in response to poor DSST
results. Gupta et al. (2008) and others advocated for a clearer diagnostic approach to
model evaluation based on hydrologic signatures, chosen to be as ‘theoretically rele-
vant’ as possible. In this view, signature replication gives insight into which aspects
of the flow regime are poorly matched, insight not possible with global performance
scores. Gupta et al. (2008) suggest to deduce which component of a model structure
relates to a signature shown to be poorly matched, and focus on that component.
de Vos et al. (2010)) split historic data into temporal clusters based on hydrologic
similarity, calibrating the same hydrologic model structure to each cluster in turn.
Their method of model improvement assumes that “deficiencies of the model struc-
ture cause the model parameters to vary ... to compensate for the effects of the
model structural error” (p2841, cf. Beck, 2002). Westra et al. (2014) used similar
logic, allowing parameters to change value with time based on selected co-variates
(Julian day, recent climatic conditions, or a linear trend) as an intermediate step
towards model improvement.

However, deficiencies in simulations may be caused by a variety of other fac-
tors besides model structural inadequacies. For example, the first step advocated
by Westra et al. (2014) is to check the data for systematic errors (eg. due to a
non-stable rating curve - eg. McMillan et al., 2010). In a study of multiple model
structures, Shoai\textit{b et al.} (2016) show how uncertainty in model simulations can be due to model structural choices, selection of objective function, and issues of parameter identifiability (and also input errors in the later work of Shoai\textit{b et al.}, 2018). Along similar lines, Coron \textit{et al.} (2014) provide a list of ‘potential explanations for robustness problems’, namely: (i) an ineffective model structure; (ii) an inappropriate calibration strategy; (iii) temporal changes in data errors; (iv) temporal changes in the catchments’ natural function; and/or (v) temporal changes in anthropogenic impact. Recent evidence in support of item (ii) include the inter-method comparisons of Chapter 4, which revealed that common ‘least-squares’ type approaches to calibration often miss parameter sets that are robust to changes in climatic conditions. Least squares approaches may give weight to model behaviour during time of high flow (eg. Freer \textit{et al.}, 1996; Legates and McCabe, 1999; Krause and Boyle, 2005), while being relatively insensitive to parts of the flow regime (eg. mid- to low-flows) which in certain cases (eg. future drying) may be a closer analogue of future conditions. Poor simulations may also arise from a lack of informative calibration data, particularly if different processes are active in the prediction period compared to the evaluation period (Reichert and Omlin, 1997; Ljung, 1998; Andr\textit{e}assian \textit{et al.}, 2012, see for example the climate induced cessation of saturation excess processes discussed by Kinal and Stoneman, 2012), or from disinformative data affecting parameter selection (Beven, 2011). Thus, in many cases, model structural change may not be the most appropriate action.

Based on the above review, a gap in the existing literature is a holistic framework that includes consideration of all of the items on Coron \textit{et al.} (2014)’s list. This chapter attempts to define such a framework. The main question to be answered is ‘When a model fails a DSST, what process of improvement should be followed?’ Some steps in the framework could be undertaken by many existing methods; we anticipate the main value of this chapter is not in the details of each step, but rather in the linking of the steps into a unified framework.

In this chapter, Section 7.3 contains a general description of the framework, and Section 7.4 outlines an application of selected steps of the framework to a case study
from Australia. Section 7.5 discusses various aspects of the framework, including the relation between DSST results and model adequacy, the potential to apply the framework for large sample hydrology (Gupta et al., 2014), and the implications of the framework for the interpretation of DSST results.

7.3 Overview of framework

The framework is shown visually in Figure 7.1, and the steps are described below. Here ‘model structure’ means the model equations (including underlying conceptualisation and process representation) and their numerical implementation; ‘model’ means a parameterised model structure (Andréassian et al., 2009, sometimes ‘parameterised model’ is used to make this clearer); and ‘evaluation period’ is used in place of ‘validation period’ (cf. Oreskes et al., 1994).

The colours in Figure 7.1 divide the steps of the framework into three categories. The first category (blue) are tasks related to the DSST, which provides an initial model evaluation. As stated, the framework is intended for use with models that have already failed the DSST, in which case the blue tasks should already have been done, either implicitly or explicitly. The red and green categories aim to diagnose where model failures are coming from (which could be calibration issues, data issues, or structural inadequacy), and to correct these if possible. The red boxes are applications that are new for this study, including the Pareto Test, which is central to the framework, and a suggested method for analysis of signature drift that informs Step 4b (described in Section 7.4.10). The green boxes denote steps that are undertaken contingent on the Pareto Test result. Analysis methods to complete these tasks are generally available in existing literature, and are thus given less focus here.

1a. Choose acceptance metrics and thresholds. Key question: what aspects of model performance are important, and how can these be quantified? Since users of this framework are expected to have already conducted a DSST, with unfavourable results, such users will usually have already done the following steps
Figure 7.1: Flow chart of the model improvement framework. The Pareto test complements the Differential Split Sample Test (DSST) and provides additional information to guide whether structural changes are required. The blue tasks are tasks normally undertaken (explicitly or implicitly) as part of the DSST, while the red tasks are new for this framework, and green tasks are undertaken contingent on step 3 (Pareto Test) results. 'DSST failure' means the parameter set(s) chosen by a given calibration method (which uses calibration period data only) are unable to meet the acceptance thresholds. In contrast, 'Pareto failure' means that no parameter set in a model structure is able to meet the acceptance thresholds, and thus structural changes or improved data are required. DSST failure does not imply model structural failure in cases where the desired performance is attainable by the model structure, but the calibration method failed to find the right parameter set(s) ('fail' in DSST, 'pass' in Pareto). With regards to the decision point at the asterisk (*), note that there is no way of knowing a priori whether information is sufficient, so a way forward is to initially seek an improved calibration method with the data at hand (ie. upper option), introducing additional data later as required/available (lower option, marked with dotted arrows).
7.3. OVERVIEW OF FRAMEWORK

(either implicitly or explicitly): (i) decided what flow behaviours are most important for the model to replicate; (ii) decided on an appropriate way of quantifying this behaviour through one or more metrics - called ‘acceptance metrics’ in this chapter; and (iii) decided - sometimes implicitly - on some threshold of acceptance (which the model has subsequently failed to meet). For example, for a flood model: (i) replicating flood events is most important; (ii) the acceptance metric might be the error in simulated discharges or volumes for selected flood events; and (iii) the threshold might be decided based on estimated uncertainty bounds on flood event data. Alternatively, for a water resource model, (i) the important flow behaviours to replicate might be the long term average flow, and the timing of runoff on timescales relevant to water infrastructure; (ii) corresponding acceptance metrics might be model bias and the NSE on daily and/or monthly timescales, and (iii) each measure would need a separate threshold reflecting the modeller’s understanding of the level of accuracy required for the decision making purpose of the model and/or uncertainty in the data. To encourage clarity of purpose, we recommend that such decisions are made prior to any modelling; hence they are placed first in this framework.

1b. Define periods. Key question: over what periods of time should model calibration and evaluation take place? A modeller that has conducted a climate-based DSST will have already defined time periods for calibration and evaluation, such that there is a contrast in climatic forcing between periods. Models are required to meet thresholds from 1a over the calibration period, in addition to the evaluation period(s).

1c. Choose calibration method. Key question: how should parameter set(s) be selected? A modeller who has conducted a climate-based DSST will have selected a calibration algorithm, the purpose of which is to search the parameter space for parameter sets with desirable behaviour in one or more objective functions. This step encompasses the choice of both the algorithm and the calibration objective function(s). Often, the default choice for objective function is to use the acceptance metric chosen in Step 1a (or, if multiple, a weighted sum). Nonetheless, a clear distinction is useful between the acceptance metric(s) (used to characterise
whether a model has attained an acceptable level of performance) and objective function metric(s) (used to select parameter sets) - even if, in fact, the same metric or metrics are used for both purposes. This distinction allows for the trial of different calibration methods (including multi-objective methods) while the overall purpose of the model, as expressed in the acceptance metric(s), remains unchanged. Although it may be argued that the best chance of maximising a certain metric in the evaluation period is to optimise the same metric in the calibration period, this is not universally true: robust simulation performance in drier or wetter conditions than the calibration data depends on fidelity of process representation, so calibration objective function(s) should be chosen to extract information relevant to these processes from the calibration data (an example is given in Section 7.4.14).

Since users of this framework are expected to have already conducted a DSST, with unfavourable results, such users will have already conducted Step 1c, determining that the ‘optimum’ parameter set identified by their calibration method does not meet the acceptance criteria; or alternatively, if using ensemble methods (ie. those that identify multiple parameter sets rather than just a single ‘optimum’), they will have determined that the ensemble does not contain parameter sets that meet the acceptance criteria, or perhaps that an insufficient proportion of the ensemble meets the criteria. It is expected that users will have checked the reliability of their method, eg. to ensure that it is not reporting a secondary optimum.

2a. Model calibration. In this step, model calibration is implemented using the preselected method. This step may only use information in the calibration period, by definition (and in contrast to Step 3).

2b. Calculate acceptance metric values. Models that meet acceptance thresholds are what Klemeš (1986) referred to as ‘operationally adequate’ in the sense that they have met a set of criteria that are deemed relevant (Step 1a) to the operational purpose of the model (while not necessarily obtaining these answers for the right hydrological reasons - more on this in Section 7.5.3). In contrast, the case where acceptance thresholds are not met (DSST failure), is the intended context of the remainder of the framework. The steps up to this point give an initial model
7.3. OVERVIEW OF FRAMEWORK

evaluation using the DSST. Steps 3 and onwards aim to diagnose where model failures are coming from, and to correct these if possible.

3. Pareto Test. **Key question:** does any parameter set in the model structure fulfil all acceptance thresholds? The Pareto test complements the DSST and provides additional information to guide whether structural changes are required. Consider the following two statements:

a. The parameter set(s) chosen in the DSST failed to meet acceptance thresholds.

b. All parameter sets in the model structure failed to meet acceptance thresholds.

The preceding DSST-related steps tested statement (a). In contrast, Step 3 tests whether (b) is true, via a multi-objective calibration that directly calibrates the model simultaneously to all acceptance metrics, including those defined over the evaluation period(s). The output, a Pareto Front, is a set of ‘non-dominated’ parameter sets - sets for which it is not possible to obtain a higher score in one acceptance metric without worsening the score(s) of the other acceptance metric(s) (Efstathiadis and Koutsoyiannis, 2010). Since some acceptance metrics are defined over the evaluation period(s), the Pareto Test uses information in the evaluation period, so that the evaluation period is no longer independent (but the phrase ‘evaluation period’ continues to be used in this chapter, for clarity). This is permitted only because it is intended as a diagnostic tool, not a calibration tool (cf. Chapter 4). Any Pareto-based multi-objective calibration algorithm may be used for Step 3. The Pareto test gives the ability to differentiate between two cases of model failure. A ‘fail’ in the Pareto test means no parameter set in the model structure can fulfil all acceptance thresholds, no matter how it is calibrated. This indicates the need for model structural change or data quality review (Step 4), assuming the Pareto algorithm works as designed (cf. Kollat et al., 2012). A ‘pass’ in the Pareto test means the acceptance thresholds are attainable by the model structure, even though the calibration step of the DSST failed to find parameter set(s) that meet the thresholds. In case (b) the option is available to focus on improving the calibration method (Step 5).
4a. Analyse change. Key question: are there any temporal changes in data errors, human impacts or natural processes that could help to explain the Pareto failure? Following the list of Coron et al. (2014), this step involves the following tasks. Firstly, examine the data (Coron et al.’s item iii) to determine whether error structures are changing with time in a way that could explain the Pareto failure (eg. due to a non-stable rating curve - eg. McMillan et al., 2010; or due to developments or declines in the monitoring network). Secondly, assess whether human impacts (Coron et al.’s item v) have changed with time: this could include water pumping, impoundments such as small on-farm storages, or any other human activity that could intercept or add water to the system. Thirdly, examine evidence of changes in natural processes (Coron et al.’s item iv) that could explain the failure of the model at Step 3, possibly seeking data such as groundwater levels, soil moisture, snowpack, or vegetation data (Section 7.4.9). This data may be scarce, possibly restricting activities in this step. As Step 4b’s characterisation of model failure informs Step 4a, 4a and 4b should be done concurrently.

4b. Analyse model structure. Key question: based on analysis of deficiencies in simulations and sensitivity analysis, which part(s) of the model structure should be altered? The goal is to narrow down the problem, if possible, to a particular component of the model structure. This may include tests of whether optimal (or near-optimal) parameters are changing with time, on short and long timescales, for which multiple methods are possible (eg. Wagener, 2003; Marshall et al., 2007; de Vos et al., 2010; Merz et al., 2011; Westra et al., 2014). A complementary analysis (presented in Section 7.4.10) is to determine aspects of the flow regime for which the quality of simulation ‘drifts’ systematically with climate, through an analysis of hydrologic signatures (cf. Gupta et al., 2008), and then relate this to model parameters via sensitivity analysis.

4c. Model structural or data improvement. Based on the results of Steps 4a and 4b, make changes to the model structure. Alternatively, if a strategy of allowing parameter values to vary with climatic conditions is preferred (eg. Oudin et al., 2006; de Vos et al., 2010), implement this strategy; or if the problem was due
to data errors, remedy these errors if possible. Return to Step 2.

If the model structure passes the Pareto Test, this opens up Steps 5a and 5b (note, continued model structural improvement (4a-c) also remains an option, although this is not marked on Figure 7.1). Any effort to improve calibration methods should bear in mind that there may not be sufficient information in the calibration data to parameterise the model. This could be due to processes that are inactive or unimportant during the calibration period that become important when climatic conditions change. If so, it may be unreasonable to assume that the model parameters that govern such processes would be identifiable based on pre-change data alone (Reichert and Omlin, 1997; Ljung, 1998; Andréassian et al., 2012; see van Werkhoven et al., 2008 for a demonstration of this issue across space). Alternatively, information regarding these processes may be present in the calibration period in non-discharge data types (eg. groundwater, soil moisture, remote sensing), so that calibration strategies that use this data may be favoured. However, this is difficult to know a priori, so a way forward is to initially seek an improved calibration method with the data at hand (Step 5a), introducing additional data later as required and if available (Step 5b).

5a. Revise calibration method. Key question: can a different calibration method give improved DSST results? This option seeks an improved calibration method for the DSST, with no addition of data. In choosing whether to undertake this step, modellers face a dilemma. Because this step trials at least one additional calibration method, it involves entering a feedback loop that chooses the calibration method based (at least partially) on model performance in the evaluation period, thus compromising strict adherence to the idea of independence in split sample testing. On the other hand, not undertaking Step 5a when the model structure has already passed the Pareto Test, means knowing the model structure is capable but not allowing oneself to use this capability. This dilemma should be a strong motivating factor for a well-informed initial choice (Step 1c) of calibration method. For those proceeding with this step, the literature on calibration methods is expansive, and potential options include changes in objective function, changes in calibration algorithm, multi-
objective methods including sub-period analysis, and limits of acceptability methods (cf. Section 7.5.2). To minimise the sense of ‘trial and error’, methods should be rationally chosen with recourse to hydrological theory (i.e. why would we expect an improved outcome, given the purpose of the model?).

5b. Additional data. Key question: does a calibration strategy that incorporates additional data lead to improved DSST results? Repeated cycling via Step 5a without successfully passing the DSST may mean that, although the model structure contains parameter sets that meet all objectives, there is insufficient information in the calibration period data to identify them uniquely, in which case Step 5b may assist. However, this step may be undertaken for other reasons, such as increasing the realism of the model (Clark et al., 2011). Additional data could include observations of groundwater, soil moisture, vegetation data, or snow data, in addition to qualitative or fuzzy measures that reflect subjective understanding of dominant processes.

This ends the general description of the framework.

7.4 Case study

The purpose of this section is to apply the framework to an example catchment. For brevity, many steps contain a summary only with further information in Appendix E.

7.4.1 Example catchment and data

The example takes the perspective of a water-resources focussed climate change impact assessment. Thus, it aims to develop a rainfall-runoff model suitable for water resources assessments under projected future climatic conditions. Given this context, an example catchment is selected that contributes to a metropolitan water supply, with a history of significant seasonal and interannual variability and with good quality records, all of which make it an interesting and suitable study for demonstrating
the framework. The selected catchment is Harvey River in the south west of Australia (Figure 7.2a), upstream of the gauge at Dingo Road (station 613002, area 148 km², mean annual rainfall 1000 mm/yr, runoff ratio 0.2). Downstream of the gauge, the river flows into the 57 GL Stirling Reservoir which supplies water to numerous communities including the city of Perth. The catchment has a strongly seasonal climate with hot dry summers and cool, wet (but snow-free) winters, with >80% of flow occurring in the months of July to November. The catchment is entirely forested, and is dominated by Jarrah hardwoods (Eucalyptus marginata). The gauge is one of 222 ‘Hydrologic Reference Stations’ selected by Australia’s Bureau of Meteorology as having “minimal water resource development and land use disturbances” (Turner, 2012, p6). Year-to-year variability is high (Figure 7.2a), and a run of low-rainfall years in the 2000s led to a cumulative rainfall deficit, relative to earlier decades. The associated region-wide decline in streamflow led to construction of a desalination plant for Perth and a greater dependence on groundwater (Petrone et al., 2010).

A lumped modelling approach with a daily timestep is adopted, with daily catchment average rainfall derived from the interpolated (5km) gridded rainfall product of Jones et al. (2009), and Potential Evapotranspiration (PET) estimates
derived according to the Wet Environment method from Morton (1983), using the
gridded dataset produced by Jeffrey et al. (2001) extracted at the catchment centroid.
The extracted PET has an average annual value of 1340 mm/year, approximately
35% greater than rainfall. Catchment boundaries are derived using D8 flow analysis
on post-processed Shuttle Radar Topography Mission (SRTM) data published by
Gallant et al. (2011), on a grid size of 1 second (approximately thirty metres).

7.4.2 Example model structures

Since the focus is on the framework itself, this chapter presents no new model struc-
tures. Instead, we choose a model structure which has two pre-existing variants that
provide a case study in model improvement: namely, IHACRES. Here, both versions
have a daily timestep and are spatially lumped.

The first version, termed IHACRES-A in this chapter, follows the descriptions
of Jakeman et al. (1990) and Jakeman and Hornberger (1993) with six free param-
ters. IHACRES-A implements an index of catchment wetness that increases linearly
in response to rainfall (as a function of parameter c), and decreases non-linearly in
response to PET (as a function of parameters c, Tw and f). The index value de-
termines the proportion of rainfall converted to runoff. Runoff is routed through
two parallel linear stores with different time constants (parameters Tq and Ts). The
split between the routing stores is determined by parameter Vs. Parameters and
thresholds are provided in the Appendix, Table E.1.

IHACRES-B includes changes by Ye et al. (1997), who retained the index
of catchment wetness but allowed for a non-linear relationship (described by new
parameter p, pâèl) between the index and runoff. They also enforced a threshold
on index values (set by new parameter l, l âè0) which must be exceeded before any
rainfall can be converted to runoff. Since they worked in ephemeral catchments, Ye
et al. (1997) removed the slower of the two routing storages. However, given Harvey
River runoff is relatively sustained and historically perennial, it is appropriate to
retain both routing storages in IHACRES-B, giving eight free parameters. Thus,
IHACRES-B is an extension of IHACRES-A and IHACRES-A is a special case of
IHACRES-B with \( p = 1 \) and \( l = 0 \). Note that IHACRES-B is the same model as used in all previous chapters, referred to previously as simply ‘IHACRES’.

### 7.4.3 Application of framework

In the following demonstration of the framework, we begin with IHACRES-A, moving through steps 1-4. Step 4b demonstrates that the additional parameters added by Ye et al. (1997) are supported by the data for this case study. Using the additional parameters as an example of model structural improvement, we swap to IHACRES-B for Step 4c and subsequent steps.

### 7.4.4 Choose acceptance metrics and thresholds (Step 1a)

*Key question: what aspects of model performance are important, and how can these be quantified?* Given the water resources context of the case study, it is important to replicate runoff volumes, at timescales relevant to the water supply system response (days-weeks). We adopt the Kling Gupta Efficiency (KGE - Gupta et al., 2009) to quantify this. KGE is a quasi-least-squares metric that quantifies the match in mean flow, variability in flow, and timing (via linear correlation). For the purposes of illustration, the acceptance threshold is set to \( \text{KGE} = 0.8 \), which ensures that the mean and variability will not be biased by more than 20\%, nor will the correlation be less than 0.8.

### 7.4.5 Define periods (Step 1b)

*Key question: over what periods of time should model calibration and evaluation take place?* According to climate projection studies summarised by Whetton et al. (2016), future climatic conditions in the south west of Australia are likely to be drier than in the past. Thus, to match the test to the intended application as much as possible, the evaluation period is chosen to be drier than the calibration period (Figure 7.2). The evaluation period is defined as the seven driest consecutive years in the record, which are 2005 to 2011, as shown in Figure 7.2. The calibration period is the preceding years 1970 to 2004. Thus, the evaluation period acceptance threshold
is $KGE_{2005-2011} \leq 0.8$. A separate acceptance threshold is set for the calibration period, $KGE_{1970-2004} \leq 0.8$.

7.4.6 Choose calibration method (Step 1c)

*Key question: how should parameter set(s) be selected?* For the purposes of illustration, we initially set the calibration objective function as the same metric as the acceptance metric - KGE. The KGE over the calibration period is optimised using the evolutionary single objective optimiser CMA-ES (Hansen, 2006), which was found by Arsenault et al. (2014) to perform favourably compared to other algorithms commonly used in hydrology. CMA-ES settings and other details are outlined in Section 6.3.3. Although single objective optimisation is not the most insightful calibration method - for example, ensemble methods may give a greater understanding of the capabilities of the model - it is the most popular method in the literature and in practice, so for the purposes of illustration it is useful to start with single-objective optimisation and show what lessons may be learned through application of this framework. Further, the KGE may not be the most appropriate objective function for use in a drying climate (more on this in Section 7.5.2), but it is a useful starting point here as an illustration, particularly to demonstrate the step to improve calibration methods (5a). Step 5a discusses principles of selection of calibration methods further, and in practice these should be applied from the outset.

7.4.7 Model calibration and calculation of acceptance metric values (Step 2a & 2b)

As noted, in this case study we begin with IHACRES-A. Running the calibration with the preselected method gives the following scores for the acceptance metrics for IHACRES-A:

- $KGE_{1970-2004} = 0.91$
- $KGE_{2005-2011} = 0.45$
Thus, the KGE over the evaluation period is well below the acceptance threshold, and IHACRES-A fails the DSST. The combination of high calibration score with low evaluation (post-change) score is consistent with the literature review (eg. Vaze et al., 2010).

7.4.8 Pareto test (Step 3)

Key question: does any parameter set in the model structure fulfil all acceptance thresholds? In contrast to the previous step, the Pareto Test directly calibrates the model to both acceptance metrics simultaneously, so the calibration algorithm must have access to the 1970-2004 data and the 2005-2011 data (ie. 2005-2011 is no longer independent in the Pareto Test). The evolutionary multi-objective optimiser AMALGAM (Vrugt and Robinson, 2007) is used to define a Pareto Front, using a population of 100 parameter sets (Section 4.3.8). AMALGAM was shown by Vrugt and Robinson (2007) to perform efficiently compared to similar algorithms used in the literature.

The Pareto curve results are shown in Figure 7.3. Parameter sets at the left hand end of the Pareto curve meet the first acceptance threshold but not the second (cf. the DSST results of the previous section), and parameter sets at the right hand end, vice versa (ie. the model performs well over 2005-2011 provided it is calibrated to it in isolation). However, no part of the Pareto Curve passes through the space where both thresholds are met, indicating a lack of robust parameter sets in IHACRES-A. The implication is that IHACRES-A is incapable of meeting both acceptance thresholds with a single parameter set, no matter how it is calibrated. In such situations, there is little use attempting to improve the DSST calibration method. Thus, we instead review the model structure and data quality.

7.4.9 Analyse change (Step 4a)

Key question: are there any temporal changes in data errors, human impacts or natural processes that could help to explain the Pareto failure?

For brevity, this section contains only a summary, but more detail is provided
in Appendix E.3. Note also that Table E.2 is a reference list of possible causes of change (in general, not specific to this case study) for each of the three categories (data errors, human impacts, natural processes).

Firstly, to provide context for assessing changes in other variables, we provide summary statistics of the calibration and evaluation periods. Estimated mean annual values, in millimetres per year, are:

- 1974-2004: precip.: 1003; PET: 1340; runoff: 224 (recorded) & 221 (IHAC-A)

- 2005-2011: precip.: 936; PET: 1360; runoff: 133 (recorded) & 192 (IHAC-A)

The quoted IHACRES-A values correspond to simulated runoff using the parameter set from Step 2a. Thus, relative to the 1974-2004, 2005-2011 saw a 7% reduction in rainfall, 1.5% increase in PET, and 40% reduction in observed runoff.
IHACRES-A overestimated runoff in 2005-2011 by 44%.

As explained in Appendix E.3, we find little evidence of temporal changes in data errors or human impacts that could explain simulation discrepancies of this magnitude. There is no evidence of temporal changes in errors in rainfall or runoff data. Regarding PET, it is possible that locally increasing trends in wind (cf. Donohue et al., 2009, p27 - note that this contrasts with more general results reported by McVicar et al., 2012) may have caused changes in evaporative demand that are not characterised in the adopted PET formulation (the ‘Wet Environment’ method of Morton, 1983, which does not require wind data). Being an undeveloped and entirely forested catchment, it is unlikely that changes in human practices are impacting on flow. In terms of natural processes, it is unlikely that wild fires caused large impacts on recorded flow, based on a review of fire history, but it is difficult to assess changes in other natural processes, eg., dynamics of vegetation and groundwater, due to lack of data.

Although there is little evidence of temporal changes in data errors, human impacts or natural processes for this catchment, the following actions could hypothetically have been undertaken in this step: (i) in the case of a non-stable rating curve, restrict analysis to portions of the historic record with a stable curve (cf. McMillan et al., 2010); (ii) in the case of un-accounted-for impoundments or extractions, estimate resulting reductions in flow over the historic record, and add them back to runoff (cf. Fowler et al., 2016); (iii) in the case of wildfire, apply a mathematical model to account for regrowth of vegetation (eg. Kuczera, 1987); (iv) in the case of non-stationarity in groundwater levels, allow this to inform how the model structure is altered in Step 4c.

7.4.10 Analyse model structure (Step 4b)

Key question: based on analysis of deficiencies in simulations and sensitivity analysis, which part of the model structure should be altered?

A wide variety of sensitivity analysis and other techniques may be relevant to this question, and those presented here are intended as a sample and guide only. The
interested reader is referred to studies and reviews such as Beck (1987, 2002, 2005); Saltelli et al. (2006); Tang et al. (2007); Marshall et al. (2007); de Vos et al. (2010); Bennett et al. (2013); Westra et al. (2014); Pianosi et al. (2015) and Pianosi et al. (2016). This section is arranged as follows: firstly, a new signature-based method is applied and the results shown. Secondly, a summary of two additional methods is provided (described in full in Appendix E.3) giving additional (and consistent) evidence in support of structural change. A user of the framework may ask whether multiple methods are strictly required. Although one method may be sufficient to form a way forward, we concur with Pianosi et al. (2015) who recommend that good practice includes “the application of multiple ... methods ... to complement and validate individual results” (p80).

Signature ‘drift’ analysis

The Pareto results (Step 3) show that IHACRES-A is unable to meet the acceptance thresholds over the contrasting periods, no matter how it is calibrated. However, given the KGE is a global metric, the Pareto results can’t provide specifics about what is going wrong. Since hydrologic signatures can characterise different aspects of the flow regime, signature analysis provides more specific insight as to how the simulations diverge from reality, and how this changes with climate. This section demonstrates a new method combining the idea of signatures as diagnostic tools (eg. Gupta et al., 2008) with ideas regarding multiple (or sub-) period analysis (eg. Coron et al., 2012; Thirel et al., 2015b). The basic idea is to identify signature(s) for which the error ‘drifts’ when climatic boundary conditions change. The steps, demonstrated visually in Figure 7.4, are:

1 Define ensemble of parameter sets for testing. Here, the 100 sets in the Pareto curve are used, but other applications might use alternative methods of generating an ensemble.

2 Divide historical record into short segments. The record for 1970-2011 is divided into eight segments of five years length. This duration is chosen
because it is long enough for signatures to reflect mid-to-long-term averages, but short enough for there to be a reasonable sample size (8 periods). Note, these periods are not related to the previous calibration/evaluation periods.

3 **Plot signature error against climate for each segment, for one parameter set.** This is done separately for each signature. Given the example is water limited, climatic conditions are quantified by mean annual rainfall, but other metrics could be used (PET, temperature, aridity).

4 **Calculate summary statistics.** We adopt two statistics: Spearman Correlation ($\rho$), which quantifies strength of relationship between climate and degree of signature matching; and the average signature error across all periods, which quantifies persistent (possibly climate-insensitive) bias in signatures.

5 **Repeat for other parameter sets.**

6 **Prepare summary statistics across all parameter sets.** The 100 values of both statistics are compiled and presented as boxplots.

7 **Identify signature(s) most sensitive to climate or most subject to bias.**

8 **Conduct sensitivity analysis.** For each signature from (7), the aim is to identify which parameters the signature value is sensitive to. Variance-based sensitivity analysis (VBSA) is undertaken using the method of Sobol (2001) as implemented in the SAFE toolkit (Pianosi et al., 2016).

To demonstrate the method, six signatures are chosen, representing different aspects of the flow regime. The overall flow volume is measured by the mean daily flow and median daily flow, the daily variability by the coefficient of variation ($C_v$), the ‘flashiness’ by the slow flow index, the degree to which flow is sustained into dry periods by the recession constant, and the amount of water required to ‘wet up’ the catchment following the dry season, by the wetting up index. Further signatures could be included, if required, to capture other aspects of the flow regime. Definitions and references for signatures are provided in Table 5.2.
Figure 7.4: (a) Analysis of climatic ‘drift’ in hydrologic signature error to diagnose model structural problems. Results show the error in median daily flow has the greatest sensitivity to climate (relative to other signatures) and the highest relative error. The whiskers extend a maximum of 1.5 times the interquartile range; values beyond the whisker are marked as outliers and are denoted as +. (b) Variance based sensitivity analysis based on Sobol’s method, linking the most sensitive signature from part a (the median) to model parameters. The box indicates 95% confidence intervals in the sensitivity. Sobol results indicate the median flow is most sensitive to the c parameter.

The results are shown in Figure 7.4a (Steps 1-7) and 4b (Step 8). Across all sets in the Pareto Front (Step 6), there is high climate sensitivity for error in mean, median and slow flow index, while error in \( C_v \) and recession constant are relatively less sensitive to climate. The wetting up index has a climate-insensitive bias of around 20% across all parameter sets, meaning that IHACRES-A consistently overestimates flow in the early wet season regardless of whether the climate segment is wet or dry. It is noted that these results are sensitive both to the method for generating the ensemble, and to the summary metrics used.

At Step 7 the median flow is selected as it has the highest climate sensitivity of
any signature, in addition to the highest average bias at Step 6. The Sobol results for Step 8 (Figure 7.4b), indicate that the median is most sensitive to the c parameter. This is true for both the main effects (which quantify “the direct contribution to the output variance from individual inputs [parameters]”) and the total effects (which “measure the overall contribution from an input [parameter] considering its direct effect and its interactions with all the other inputs [parameters]” (definitions from Pianosi et al., 2016). Similar Sobol results were obtained for the mean (not shown).

Thus, the signature whose error ‘drifts’ the most with change in climatic boundary conditions is the median daily flow, and this signature is most sensitive to the c parameter value, which governs the soil moisture storage capacity. The hydrological implications of this are discussed below (see subtitle ‘Synthesis’), but first we briefly consider confirmatory evidence from two other methods.

Other methods

Additional insight is provided by methods that allow parameters to change with time. Parameters changing with time may indicate the need for structural change in the associated part of the model structure (Westra et al., 2014; de Vos et al., 2010; Beck, 2002). Following Merz et al. (2011) among others, Figure E.2 shows the degree to which each parameter changes when calibrated separately to the 1970-2004 period and the 2005-2011 period. Parameters that are relatively static include Ts and Vs (routing parameters), while Tw and f (parameters governing ET) are poorly identified regardless of period. Consistent with the previous results, parameter c changes the most between the periods, taking a reduced value during the drier period.

Although the focus is on long-term dynamics, analysis of seasonal changes may also provide clues to diagnose deficiencies (e.g. Freer et al., 2003). For example, the end of summer in a Mediterranean climate tends to be the driest time of year, and analysing whether the transition from dry to wet is well simulated on seasonal scales may be relevant to diagnosing model response to longer term change in climate. Figure E.2 shows an example using Dynamic Identifiability Analysis (DYNIA - Wagener, 2003). Like the long-timescale analysis above, the results show that the c parameter
changes the most with seasonal cycle. Particularly notable are the difficulties at the 
start of the wet season (which was mentioned in the discussion of the wetting up in-
dex, above), and that the best \( c \) value at this time is lower than during the rest of the 
wet season. Note that DYNIA is not limited to shorter (event/seasonal) timescales 
and could provide analyses at longer timescales also, to analyse the sensitivity of 
signatures to parameter values (similar to Sobol in the previous section).

**Synthesis**

Together, these analyses build a consistent picture: IHACRES-A model improve-
ment should focus on the component of the model that contains the \( c \) parameter. 
\( c \) changed most with time in the long-term change test, and simulation deficiencies 
that ‘drifted’ with climate were shown to be more sensitive to \( c \) than to any other 
parameter, given the parameter ranges chosen (Table E.1 in the Appendix). What 
are the hydrological implications of these results? As per Section 7.4.2, IHACRES-A 
contains a state variable, the index of catchment wetness, which linearly determines 
the proportion of incoming rainfall converted to runoff. Precipitation causes the 
index of catchment wetness to increase and ET causes decrease. The parameter 
\( c \) specifies the proportional increase in index value for each unit of precipitation. 
Thus, the reciprocal of \( c \) has a similar function to the soil moisture store capacity 
in other models. The overestimation of flow during drier 5-year segments (Figure 
7.4) suggests a deficiency in existing relationships governing the dynamics of the 
wetness index and runoff generation. Either the relationship between the index and 
runoff generation needs changing (eg. from linear to non-linear) to achieve a steeper 
decline in runoff generation as wetness decreases, or the wetness index formulation 
needs changing so that the index attains lower values during drought. The observed 
non-stationarity of \( c \) is interpreted as follows: to dynamically fit drier periods, runoff 
overestimation is avoided by decreasing \( c \) (the equivalent of increasing the soil mois-
ture storage capacity) to absorb the excess. The final factor to consider in model 
structural change is the tendency of IHACRES-A to overestimate flow at the be-
ginning of the wet season regardless of whether recent conditions were wet or dry.
This was indicated by the signature analysis (cf. underestimated wetting up index in Figure 7.4) and corroborated by the seasonal DYNIA analysis.

7.4.11 Model structural change (Step 4c)

As noted earlier, no new model structures are introduced in this chapter. However, the deficiencies noted are closely related to the new parameters introduced by Ye et al. (1997):

a New parameter \( p \) describing non-linearity of relationship between catchment wetness index and runoff. This relationship was previously linear in IHACRES-A. This non-linearity is intended to result in reduced simulated flow in dry years and increased simulated flow in wet years.

b New parameter \( l \) enforcing a threshold on catchment wetness values. The threshold must be exceeded before any rainfall can be converted to runoff, affecting the early wet season. Thus, these additional parameters are used as an example of model improvement, and we henceforth swap from IHACRES-A to IHACRES-B. It is noted that changing the model to IHACRES-B does not resolve the hypothesis suggested in Section 7.4.9 that trends in wind caused increases in evaporative demand not characterised in the adopted PET data. For the purpose of illustration we elect to test the structural changes above, rather than pursuing the PET hypothesis further (e.g. adopting a formulation that uses wind; cf. Donohue et al., 2010; McMahon et al., 2013; Guo et al., 2017a,b).

7.4.12 Repeat model calibration and calculation of acceptance metrics (Step 2a & 2b)

This step is a repeat of Section 7.4.7 using identical methods except that the model structure has changed from IHACRES-A to IHACRES-B. The results are:

- \( \text{KGE}_{1970-2004} \) (calibration) = 0.93
Figure 7.5: (a) Pareto test results for IHACRES-A (Section 7.4.8; Figure 7.3) and IHACRES-B (Section 7.4.13). Parameter sets that fulfil both objectives exist in IHACRES-B but not IHACRES-A. Arrows show the proposed summary metric (closest distance to perfect point). (b) Repeat of Figure 7.4a results, shown for IHACRES-A (blue) and IHACRES-B (red). The sample of values (N=100) is derived from analysing each of 100 parameter sets in the Pareto curve. IHACRES-B has less climate sensitivity in error of mean and median, and less bias in Wetting Up Index. The whiskers extend a maximum of 1.5 times the interquartile range; values beyond the whisker are marked as outliers and are denoted as +.

- KGE$_{2005-2011}$ (evaluation) = 0.54

Despite the structural changes, the KGE over the evaluation period is still well below the acceptance threshold. IHACRES-B thus fails the DSST.

7.4.13 Repeat Pareto test (Step 3)

Key question: does any parameter set in the model structure fulfil all objectives?

This step is identical to Section 7.4.8 except that the model structure has changed from IHACRES-A to IHACRES-B. The resulting Pareto curve (Figure 7.5a) passes through the space where both acceptance thresholds are met. Thus, despite the poor DSST result, IHACRES-B does contain parameter sets that can meet both acceptance thresholds at the same time. Furthermore, the dynamics of the catchment wetness index for such parameter sets (Figure E.4) are more sensitive to small changes in rainfall on timescales of months to years, consistent with perceptual understanding.
7.4. CASE STUDY

In addition to repeating the Pareto test, we repeat the signature analysis (cf. Figure 7.4a) to indicate which aspects of the flow regime have improved due to the structural changes (Figure 7.5b). IHACRES-B simulations have significantly less climate sensitivity in error of mean and median, and the bias in Wetting Up Index has reduced, all in line with intentions. There remains room for improvement across all signatures, including slow flow index and recession constant which appear to have degraded slightly because of the model structural changes. Thus, a ‘pass’ in the Pareto test does not preclude further model structural improvements (ie. a return to Step 4). However, in the next section we demonstrate Step 5 (improvement of DSST calibration method) rather than repeating Step 4.

It is noted that the Pareto curves provide a way to quantify model structural improvement with respect to the acceptance metrics. Model structural performance is best quantified by how close the Pareto curve gets to ‘perfect’ scores in all acceptance metrics (black circle in Figure 7.5a). For IHACRES-A this point (filled blue circle in Figure 7.5a) is 0.35 units from the perfect point (arrow A), which reduces to 0.21 for IHACRES-B (arrow B). These numbers are better measures of model structural performance than DSST results, because the latter are sensitive to DSST calibration method.

7.4.14 Revise calibration methods and repeat DSST (Step 5a)

Key question: can a different calibration method give improved DSST results?

As mentioned in Section 7.3, this step is subject to the dilemma of whether to allow the information in the evaluation period to inform the choice of calibration method, which would violate the idea of independence in split sample testing. The guiding aim for method choice should be to extract information relevant to process representation, with reference to the model purpose. We strongly recommend that modellers consider this at Step 1c, to avoid the present dilemma at 5a if possible.

Many different calibration methods exist and it is beyond the scope of this chapter to test all methods that may be relevant to the present example. Nonetheless, we offer the following short discussion with associated testing. The present
calibration method is to maximise the KGE. The components of the KGE are most sensitive to high flow behaviour (eg. Freer et al., 1996; Legates and McCabe, 1999; Krause and Boyle, 2005) and can be strongly influenced by gauging error during a small number of timesteps during large floods (Berthet et al., 2010). Improvement may come through changing the objective function to extract information about other aspects of the flow regime, ie. mid-low flows (while not ignoring high flows) particularly since these aspects are a closer analogue of the drier conditions in the evaluation period. One way of doing this is to focus on the sum-of-absolute errors rather than the sum of squares. A further option is to explicitly ensure that years of lower flow in the calibration period are given the same weight as wetter years. Chapters 5 and 6 achieved this through a ‘Split’ KGE where each year has a separate KGE calculated, and the objective function is the average of the values from each year. This latter method is a variant on the ‘sub-period’ logic demonstrated by Freer et al. (2003); Choi and Beven (2007); Gharari et al. (2013), whereby parameter sets are assessed on their performance within fragments of the total calibration period.

In line with this discussion, the Refined Index of Agreement (based on sum of absolute error - Willmott et al., 2012) and the Split KGE, are tested in Figure 7.6. It is seen that both methods result in better KGE values in the evaluation period, but neither method provides simulations that meet both acceptance thresholds. Thus, we are left with the knowledge that the model structure can meet the acceptance thresholds, but we do not yet have a calibration method that can identify the behavioural parameter sets using 1970-2004 data alone. This provides a vivid example for the need for continued research and systematic testing of model calibration methods (as discussed further in Section 7.5.22).

7.4.15 Additional data (Step 5b)

Key question: does additional data lead to improved DSST results? For brevity, this step is not undertaken in this chapter. Broad guidance on the kinds of tasks this step may involve, is provided in the Appendix E.4.
7.4. CASE STUDY

Figure 7.6: Comparison of DSST results for different DSST calibration methods. In each case the model is calibrated to 1970-2004 and evaluated over 2005-2011. Both alternative objective functions outperformed the original calibration method (optimisation to KGE), although neither met both the acceptance thresholds.

7.4.16 Summary of case study application

To summarise, this case study began with IHACRES-A failing the DSST. Subsequent Pareto testing indicated that no parameter set could meet the acceptance thresholds, which led to data review, analysis of model structure, and subsequent structural changes (by swapping to IHACRES-B). IHACRES-B also failed the DSST, but the Pareto test indicated that there are parameter sets in IHACRES-B that meet the acceptance thresholds. Alternative calibration methods yielded some improvement, but further research is required to identify a calibration method capable of identifying parameter sets that exceed both acceptability thresholds.
CHAPTER 7. MODELLING IMPROVEMENT FRAMEWORK

7.5 Discussion

7.5.1 Interpretation of the DSST

As already emphasised, this thesis makes a clear distinction between different types of model failure, and we recommend that this distinction be carried on in future practice. DSST failure (the failure of a parameterised model to fulfil acceptance thresholds) was contrasted with Pareto failure (the failure of every parameter set in a model structure to fulfil acceptance thresholds, due to model structural failure or data errors). In the literature, separate tests to distinguish these error types (such as Step 3) are uncommon, and DSST failures are often interpreted as implying both model failure and model structural failure. This can be profoundly misleading. For example, consider possible interpretations from the following DSST results (cf. Section 7.4.7 and 7.4.12):

- 1970-2004 (calibration period) KGE: IHACRES-A: 0.91; IHACRES-B: 0.93;
- 2005-2011 (evaluation period) KGE: IHACRES-A: 0.45; IHACRES-B: 0.54.

Interpreted under the false notion that DSST failure implies model structural failure, two conclusions would be: (i) both structures are unable to meet acceptance thresholds; and (ii) the structural changes (Section 3.3.8) made minimal difference to performance. As demonstrated, both conclusions are incorrect. For this reason, we strongly recommend future studies conduct targeted tests to distinguish between failure types, such as the Pareto test demonstrated at Step 3, or other tests based on similar multi-objective logic.

7.5.2 Further research in calibration methods

The case study demonstrates the importance of calibration methods in modelling outcomes, and further research is recommended in this area. As stated, choice of calibration method (including selection of objective function) should be grounded in hydrological theory and be informed by the purpose of the model. One potential problem for the practitioner is that there are many studies suggesting potential
calibration methods, but little guidance to choose between them. For example, for water-resource-focussed climate change impact studies, a practitioner is faced with many methods with strong supporting theory which have been developed or demonstrated for changing climates, including (i) those that use calibration data differently by defining wetter and drier sub-periods of the calibration period Gharari et al., 2013, see also Freer et al., 2003 and Choi and Beven, 2007); (ii) trading space for time approaches which incorporate information from other catchments in the same region to predict hydrologic response (eg. Singh et al., 2011); (iii) Use of alternative objective functions that place different focus on hydrological behaviours, eg. applying transforms on flow values prior to sum of squares calculations, or using sum of absolute errors rather than sum of squares (Section 7.4.14, Chapters 5 and Chapter 6, see also Krause and Boyle, 2005; Willmott et al., 2012). Furthermore, methods that identify ensembles that are likely to be more robust to change such as (iv) data depth approaches (Bárdossy and Singh, 2008) and (v) limits of acceptability approaches (Liu et al., 2009) may also be relevant to the context of changes in climatic conditions. We recommend three classes of study to increase the value of this existing work: (a) inter-method comparison, to guide practitioners in choosing between methods; (b) studies combining ideas from different methods - for example, methods iii and iv could be gainfully combined; and (c) studies testing calibration methods in regions with relatively high hydroclimatic variability (eg. the ‘crash tests’ of Coron et al., 2012), to provide assurance that the methods work successfully on the most challenging available data, while acknowledging that this does not guarantee adequacy for future conditions, as discussed below.

7.5.3 Are models that pass the DSST ‘adequate’?

As mentioned in the introduction, a common goal when DSSTs are applied is to assess the adequacy of a model for simulation in the future. To what extent is a model that passes the DSST ‘adequate’ for simulation under projected future climatic conditions? Below we present multiple levels where success of a model at one level is necessary but not sufficient for success at the next level (Figure 7.7). Each level
Figure 7.7: Reasons why models passing a DSST may still be inadequate to simulate runoff under projected future conditions, expressed in a hierarchical scheme.

has distinct causes of inadequacy, as discussed below.

Consider the set of all models (orange in Figure 7.7) that pass the DSST by fulfilling acceptance thresholds. No acceptance metrics can perfectly reflect the model purpose; thus, simulations from models in this set may be deficient in ways not captured by the criteria (Krause and Boyle, 2005). Various strategies may help to detect such deficiencies, including referring to a wider set of criteria (Gupta et al., 1998, 2008), using a variety of checks and visualisations (Bennett et al., 2013; Thirel et al., 2015b), and matching criteria to the context as closely as possible.

Moving to the next level in Figure 7.7, models that display an adequate match (however this is defined) in numerical outputs may not do so for the right reasons (Klement, 1986; Beven, 2006). For example, a model may activate a process that experimental data or site experience reveals is the wrong mechanism for the catchment (eg. Hortonian flow in a catchment with high infiltration). Strategies for revealing this kind of inadequacy include site familiarisation/seeking local knowledge (Holländer et al., 2014), more difficult tests through extra data (eg. Mroczkowski et al., 1997), and dialogue between modellers and experimental hydrologists (Seibert and McDonnell, 2002). All of these may help to discriminate between parameters sets that are equifinal with respect to numerical output. We note that the meaning of
‘right reasons’ may vary depending on the underlying perceptual model and philosophical viewpoint of the modeller (eg. Gupta et al., 2012).

Finally, even models that match historic data for the right reasons (green) may not be adequate to simulate under projected change (blue) because the future may be so different from the past as to change the mechanisms that govern the rainfall-runoff relationship (cf. Saft et al., 2015). New processes may become dominant, and living components of the system (eg. vegetation) may respond unexpectedly due to complex feedbacks (Curtis and Wang, 1998; Rodriguez-Iturbe et al., 1999; D’Odorico and Porporato, 2004) possibly leading to less resilience following disturbances (Peterson et al., 2009). The concept known as ‘trading space for time’ (Peel and Blöschl, 2011) may be useful, whether for model parameterisation (Singh et al., 2011) or to test parameterised models in climatic conditions beyond the study area’s historic record by using another catchment entirely (eg. proxy basin split sample tests, Klemes, 1986; Refsgaard et al., 2014). One risk with such methods is that, if catchments coevolve, then they reflect the actions of processes over centuries or millennia, in contrast to climate change which is likely to occur relatively quickly. An additional confounding factor is that records of catchment hydroclimatology are themselves subject to increasing CO₂ concentrations (Roderick et al., 2015). In general, our ability to successfully transition from the green category to the blue is limited because our knowledge of the future is fundamentally incomplete. Thus, although the DSST may be “the best possible evaluation method” (Refsgaard et al., 2014), the adequacy of models that pass the DSST is far from guaranteed.

7.5.4 Generality of the framework

It is noted that the framework can be easily adapted for multiple catchments, to serve the needs of large-sample hydrology studies (Gupta et al., 2014). For example, Appendix E.5 / Figure E.3 gives an example across multiple catchments from Australia, showing how the signature analysis and the Pareto analysis can be meaningfully summarised across large samples of catchments. Furthermore, most principles in the framework are not specific to hydrology or to the context of changing climate;
thus, the framework could be used to develop modelling methods for changes across space rather than time, for changing landuse rather than changing climate, and/or to improve modelling in other areas of the environmental sciences.

7.6 Conclusions and recommendations

This chapter has presented a framework for model improvement applicable to rainfall-runoff modelling under changing climatic conditions. While model evaluation based on climate-based Differential Split Sample Tests (DSSTs) is relatively common, there is less guidance in the literature on ‘next steps’ in the case of DSST failure. The framework provides data-based guidance of whether to focus on improvement of model structures, data quality, and/or calibration methods.

We demonstrated how tests based on Pareto optimality can be used to decide if model structural changes are required. DSST failures are often interpreted as automatically implying model structural failure, and this can be profoundly misleading, possibly leading to erroneous comparisons between model structures and incorrect assessment of model structural improvements. A further contribution of this chapter is the diagnosis of structural problems via analysis of ‘drift’ in hydrologic signature error as climatic conditions change. We demonstrated a method that combined signature temporal trend analysis with parameter sensitivity analysis to provide diagnostic insight.

Many steps in the framework can be undertaken by a multiplicity of existing methods; as such, the main value of this chapter is not in the details of each step, but rather in the linking of the steps into a unified framework. We recommend future model improvement studies follow the intent of this framework, even if the methods used differ. Upon DSST failure, model structural failure should not be assumed unless indicated by separate and targeted tests as demonstrated in this chapter.
Chapter 8

Discussion and Conclusions

This final chapter begins with an overview of the research findings, linking the chapters together (Section 8.1). The limitations of the research are discussed (Section 8.2), with reference to limitations articulated in each chapter. Section 8.3 discusses future research directions, while Section 8.4 summarises the original contributions of this research. Section 8.5 provides a concluding statement for this Thesis.

8.1 Research overview

The Literature Review in Chapter 2 identified that rainfall-runoff models often exhibit poor performance when applied in different climatic conditions to the calibration period. A key research gap identified was a lack of diagnosis of the cause of this poor performance. Existing literature often assumed that model structures themselves were invalid for use in changing climates and needed to be improved, without directly testing this assumption.

However, the results of RQ1 suggest that existing model structures are more capable under changing climatic conditions than previously thought. Commonly used calibration methods often fail to identify parameter sets that are robust. The results demonstrated that such sets are often available in existing model structures, even in cases of DSST failure. Up to one half of all DSST failures were cases where the model structure was capable of the modelling standard in question, but the calibration
method failed to identify the right parameter set. Overall, RQ1 suggested that caution is needed when interpreting the results of differential split sample testing.

Since robust parameter sets often exist in existing model structures, there is a clear need for calibration methods capable of finding them, as investigated in RQ2. The RQ1 calibration method did not qualify, because it used data both before and after a change from wetter to drier conditions. In contrast, methods tested in RQ2 used pre-change data only. The results demonstrated that hydrologic simulations in changing climate strongly depend on the choice of calibration metric. Sum-of-absolute-error global objective functions tend to select more robust parameter sets, as does equally weighting all years in the calibration data. Incorporating selected hydrologic signatures in calibration functions may assist in parameter set selection, but results suggest this to be model-specific and difficult to implement. Perhaps the most significant finding for hydrological practice was that ‘least-squares’ metrics perform very poorly relative to the other available methods, and should not be used to calibrate models for a drying climate.

Lastly, RQ3 formulated a general model improvement workflow, applicable to models failing differential split sample testing, and building on RQ1 and RQ2. The existing literature is relatively devoid of guidance for the question, in what circumstances should the focus be on improving calibration methods versus improving model structures, or alternatively on other issues such as poor data quality? RQ3 filled this gap, using tests based on Pareto optimality to decide if model structural changes are required. If so, tools were suggested to diagnose structural problems, including analysis of climatic ‘drift’ in hydrologic signature error. If not, alternative calibration methods, including those from RQ2, can be trialled. This framework is the first of its kind to cover all possible causes of poor simulation, including calibration methods. This provides modellers and practitioners with practical steps for improvement of runoff projections under changing climatic conditions.
8.2 LIMITATIONS AND ASSUMPTIONS

8.2 Limitations and assumptions

Each of the research chapters provides its own discussion of limitations. The points below provide a broader overview of limitations as applied to the thesis as a whole:

1 Geographic limitations: This research was limited to catchments in the temperate zone of Australia. Thus, the improvements to calibration methods from RQ2 may not apply outside this zone. DSST failure does occur on continents other than Australia (Refsgaard and Knudsen, 1996; Singh et al., 2011; Coron et al., 2014, eg.), but the solutions may not be the same across hydroclimatic and biogeographic zones. While the logic underlying the improved calibration methods (Section 5.3.2) would appear universal, these methods should nonetheless be tested prior to application in other environments. Catchments where snow or ice play a significant role in the hydrologic cycle were not considered in this research, and the assessment of these methods in such catchments is recommended.

2 Limitations regarding models: As per Section 1.2, this research has focussed on the model type commonly used in runoff projection studies, namely daily conceptual rainfall-runoff models. While it is likely that the lessons of this thesis also apply to physically based models, this has not been tested. The improved calibration methods of RQ2 are specific to the daily timestep and would need reformulation for application on shorter or longer timesteps, although some of the principles may apply. The applicability of results to other daily conceptual rainfall-runoff model structures is untested and should not be assumed; however, given consistency of results between the five structures tested, it is considered likely that findings are general across conceptual model structures.

3 Limitations regarding data types: RQ1 and RQ2 were limited to models and methods using only precipitation, PET and streamflow. Additional data types may help to identify more robust parameter sets (Appendix E.4), but
were beyond the scope of RQ1 and RQ2. Characterisation of model adequacy (see next point) was limited to a comparison of modelled and observed streamflow. Other environmental variables, such as groundwater or soil moisture, were not considered. The PET data used in this study were from Morton’s Wet Environment method (Section 4.3.5), and other PET formulations may yield different results (see discussion in Section 7.4.9).

4 **Characterising model adequacy:** all parts of this thesis relied upon global performance metrics, in particular the KGE, as indicators of model performance. Global performance metrics cannot capture every aspect of a match between observed and simulated flow, and tend to focus on certain behaviours at the expense of others (Section 5.2). The improved calibration methods in RQ2 were judged according to whether they delivered simulations with high KGE over two contrasting periods. Since KGE and similar least squares metrics are often used in water resource assessments, the conclusions are most suited to water resource studies. Suitability for other contexts has not been specifically considered (although some insight can be gained from Figures D.7-D.11. For future studies, supplementing global performance metrics with complementary metrics such as signatures is recommended to provide a broader characterisation of simulation accuracy (Section 2.2.4).

5 **Focus on a limited number of historic periods:** To ensure a two-dimensional Pareto Front, only two historic periods were chosen ('dry' and ‘non-dry’, Section 4.3.6). Future studies may be able to enrich their findings by considering dry periods other than the driest 7 consecutive years, for example, by using methods that use time-sliding windows for a more nuanced DSST assessment (Coron et al., 2012; Thirel et al., 2015b). Such methods were not feasible in RQ1 and RQ2 due to the large numbers of catchments and models in view.

6 **Focus on calibration objectives rather than algorithms:** Improvements in calibration methods focussed on calibrating to a more suitable metric(s) rather than using a better calibration algorithm. Only a single algorithm was
used for each context - namely, CMA-ES for single objective optimisation, AMALGAM for Pareto optimisation, and DREAM-ABC for signature-based calibration. Based on published literature (Hansen, 2006; Vrugt and Robinson, 2007; Sadegh and Vrugt, 2014; Arsenault et al., 2014) it was assumed that each algorithm was the best available and was selecting the best possible parameter set(s) given the calibration objective. If incorrect, this would limit the results.

7 Lack of consideration of uncertainty: Some methods in this thesis, particularly single objective optimisation (RQ2), are limited by their lack of consideration of uncertainty. Uncertainty arises in all rainfall-runoff studies due to structural and data errors (Section 2.3.4), and methods to characterise this uncertainty exist, even for non-standard objective functions like those suggested in RQ2 (eg. Beven and Binley, 1992). In practice, hydrological projections are rarely produced for their own sake but are input into a decision making process. It is strongly recommended that rainfall-runoff studies for decision support purposes utilise these methods to ensure decision makers are aware of uncertainties arising from hydrological modelling.

8 Tested direction of climatic change: RQ2 recommended calibration methods based on their ability to produce simulations that were robust to changes in climate. In fact, only one direction of change was tested: from wetter to drier. The opposite direction remains untested and the recommended methods may not perform equally well in this direction. It is recommended that future studies investigate this further.

9 Using the past to understand the future: As outlined in Section 7.5.3, even models that provide an excellent match with historic data over a variety of climatic conditions, and for the right hydrological reasons, may still be inadequate to simulate under projected change. As noted, the future may be so different from the past as to change the mechanisms that govern the rainfall-runoff relationship. The degree of this problem varies. Some regions are already subject to considerable inter-annual variability in hydroclimate; the hydrologic
response to such variability, particularly when dry/wet periods are persistent, informs our capacity to extrapolate to future changes, as demonstrated in this thesis. However, other regions have low historic variability and projected future changes well beyond past experience. Using the past to understand the future is more challenging in such cases, and this is a key limitation of any split sample testing. To some extent, such regions may benefit from space-for-time frameworks (see next section) and from importing hydrological lessons from other places. In many cases, methods of uncertainty analysis may fail to envelop possible future outcomes, due to our fundamental lack of knowledge about the future. Split sample testing has been characterised as the best evaluation method at our disposal (Refsgaard et al., 2014), and this research has aimed at improving split sample results, but even models that pass such tests may be negated by unforeseen future changes in processes.

8.3 Future research directions

8.3.1 Better characterisation of model deficiencies

If we hope to improve model structures, it is imperative to better characterise model deficiencies, and to link this with long-term catchment change processes. These aspects are key to RQ3’s model improvement framework (Steps 4a and 4b). Without characterising what is wrong, it is difficult to propose effective structural changes. Characterisation over large samples of catchments would inform general improvements to models, as opposed to site-specific improvements that may not generalise to other locations. Research directions could include:

1 Analysis of catchment storage dynamics. As mentioned (Section 2.4.3), groundwater hydrographs suggest that, in some locations, catchment storage has a long memory and allows the effects of multiple dry or wet years to accumulate. Research questions could include: is this true generally or only for certain catchments? What type of catchments? Can remote sensing be used to inform this? Can commonly-used model structures display the same degree of
memory as observed in nature? Understanding of long term catchment changes processes could be greatly aided by more widespread groundwater monitoring in upland regions known or suspected to be sensitive to groundwater-surface water interactions (eg. the headwaters of the Campaspe and Lockdown Rivers in Victoria; see Hughes et al. 2012 for an example of the explanatory power of groundwater data in catchments in the south west of Australia).

2 Better characterisation of deficiencies in flow hydrographs. As mentioned (Section 2.2), DSST studies typically limit their reporting to global performance metrics and bias. A more detailed characterisation of the deficiencies of simulated hydrographs may improve understanding of structural deficiencies, as demonstrated in Section 7.4.10. There are many ways to do this, including analysis of climatic drift in signature error, as was demonstrated across 86 catchments for the IHACRES model structure in Figure E.3. It is recommended this be extended to other model structures.

8.3.2 Improved model structures, using the RQ3 framework

This research provided a framework for improving models under changing climatic conditions (RQ3) and it is now recommended to apply this framework to improve model structures. Based on the Literature Review (cf. Section 3.1), research topics could include:

3 Changes to ET equations: Less simplistic (possibly less linear) relationships between ET and water storage, to reflect vegetative preferential use of water deep underground as described by Dawson and Ehleringer (1991) and others.

4 Different water deficit dynamics and limits: Lessening of the ‘hard’ limits on water deficits implied by bucket-type models, since these are not consistent with groundwater hydrographs in some catchments (see previous section). Possibly, this could mean replacing such buckets with ‘softer’ systems of negative feedback that allow ET to continue, albeit at a reducing rate as catchment storage declines.
5 **Spatial dynamics**, specifically, more explicit representation of catchments as sets of spatially distributed interconnected storages. Some existing fully distributed models may permit this behaviour, as might quasi-distributed models such as Dynamic TOPMODEL (*Bven and Freer*, 2001b). Application of such models may result in different parts of the landscape retaining different memory of past hydrometeorological conditions.

8.3.3 **Improved calibration methods**

RQ2 examined calibration methods, but there is much work still to be done.

6 **Flow behaviour analysis.** As discussed (Section 5.2), different global performance metrics focus on different flow behaviours, for example least-squares metrics tend to focus on high flow components. Given that least squares metrics performed relatively poorly, this is evidently the wrong focus. Although this research demonstrated that other metrics worked better for a drying climate, the method used didn’t explicitly examine the flow behaviours that these metrics focus on. Future research might examine this, for example using the techniques demonstrated by *Krause and Boyle* (2005), with a view to better understanding and further improving calibration methods.

7 **The transition from drier to wetter.** RQ2 only considered the transition from wetter to drier; that is, DSST evaluation was undertaken on a period drier than the calibration period, not vice versa. Replication of RQ2 methods for the transition from drier to wetter is a potential future research direction and is particularly relevant for projections in the tropics (*Trenberth*, 2011).

8 **Parameter sensitivity:** Section 7.4.10 demonstrated how model performance can be related to model parameters. Similar methods could be used to investigate which parameters are most decisive in determining DSST evaluation performance. If done systematically over many catchment, it might be possible to construct more robust calibration methods using this information.
8.3. FUTURE RESEARCH DIRECTIONS

9 **Hydrologic signature analysis:** The pattern analysis in Section 5 demonstrated how the matching of certain signatures (over the calibration period) is associated with robust performance in evaluation. Two complications arose: firstly, the signatures in question were not consistent between model structures; and secondly, converting the lessons from the pattern analysis into a successful calibration method proved difficult. Research directions could include further analysis to understand why certain signatures were associated with robust flow simulations; why this changed between model structures; and whether some other method than trialled in Section 5 might be more successful in turning this into a calibration method.

8.3.4 **Other**

Two further research questions raised in Section 3.1 remain as potential topics:

10 **Monthly models:** Some literature (Niel et al., 2003; Xu, 1999) suggests that monthly models provide more stationary parameters and more robust simulations than daily models. This could be tested for a large sample of catchments and across multiple monthly model structures.

11 **Space for time:** Space for time methods have tested favourably in the USA (Singh et al., 2011), but can they provide superior hydrologic projections compared to common single-site approaches, in regions of high interannual variability such as southern Australia?

Furthermore, given that RQ2 was only applied to a finite set of catchments and model structures:

12 **Extending to other model structures and catchments:** Future research could confirm that the findings of RQ2 also hold for other model structures commonly used for projections, and in other locations with sufficient historic hydroclimatic variability for meaningful testing.
8.4 Original research contributions

Faced with literature reports of daily rainfall-runoff models commonly failing the DSST, this research:

- Developed a method to analyse the cause of reported DSST failures, focusing on distinguishing failure due to poor calibration methods from failure due to model structural deficiency;

- Demonstrated that poor calibration methods are broadly as important as deficient model structures in explaining the DSST failures. Between one third and one half of DSST failure cases (depending on the modelling standard adopted) were due to calibration method. The generality of this conclusion was assured by testing numerous model structures (5) on a large set of catchments (86) from the region (southern and eastern Australia) that provided the majority of existing large scale DSST studies (Vaze et al., 2010; Coron et al., 2012; Li et al., 2012; Silberstein et al., 2013; Saft et al., 2016a);

- Identified two strategies for improving calibration methods for use in a drying climate. Based on broad-scale testing of a wide variety of calibration methods, this research found that objective functions based on absolute error (such as the Index of Agreement) and those that consider annual as well as daily timescales (such as the Split KGE) tend to provide more robust simulations. The Split KGE is a new objective function created for this research.

- Formulated a new framework for improvement of rainfall-runoff modelling in a changing climate. Applicable to models failing the DSST, the framework uses diagnostic tests to guide focus onto model structures, data errors and/or calibration methods, as required. This is the first framework to consider calibration methods in addition to model structures and data errors. While drawing on existing tools, e.g. for sensitivity analysis, it includes several original contributions, including the use of Pareto approaches to guide which aspect of
modelling to focus on; and the analysis of climatic drift in signature error to characterise structural inadequacy.

8.5 Concluding statement

Prior to this PhD, it was known that rainfall-runoff models tend to perform poorly in changing climates, based on studies of large samples of catchments in various parts of the world. The role of calibration methods in this problem was rarely considered; rather, multiple studies stated or implied that rainfall-runoff model structures themselves are invalid for use in changing climates and need to be improved.

As a result of the original contributions of this research, the prospects for diagnosing and fixing these problems has considerably improved. The (often significant) role of common ‘least-squares’ calibration methods as a cause of poor simulations has been demonstrated. Alternative calibration methods have been tested and shown to provide superior empirical results in split sample tests. Finally, a general framework has been proposed for diagnosing the causes of poor simulations in changing climate. This framework can be applied to guide improvement efforts in individual case studies, for example to decide between improving the calibration method and improving the model structure.

This research focussed on southern and eastern Australia, a region that has recently suffered severe (in some cases, ongoing) droughts, causing significant hardship on the Australian people and environment. One of the few positive aspects of such events is that, through studying them, we may be better prepared for future environmental variability and change, allowing lessons of global applicability to be drawn from recent Australian experience.
Appendix A

Study catchment details

Below is a table listing every Hydrologic Reference Station (Turner, 2012) and explaining why the corresponding catchment was either included or excluded from study in this Thesis. Note that the list was current as at mid-2014 (221 catchments).

Table A.1: Full list of Hydrologic Reference Station catchments with reasons for acceptance or rejection as a study catchment. Column CA lists catchment area in km²; LON lists Longitude in degrees east; LAT lists Latitude in degrees south; ST lists the state or territory of Australia that the catchment lies within; and DEC lists the decision: either accepted for study (ACC), marginally accepted (MA) or rejected (REJ).

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>102101A: Pascoe River at Fall Creek</td>
<td>630.53</td>
<td>142.98</td>
<td>12.88</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>104001A: Stewart River at Telegraph Road</td>
<td>475.43</td>
<td>143.39</td>
<td>14.17</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>105101A: Normanby River at Battle Camp</td>
<td>2313.97</td>
<td>144.84</td>
<td>15.28</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>105102A: Laura River at Conlaseem Creek</td>
<td>1326.75</td>
<td>144.48</td>
<td>15.61</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
</tbody>
</table>
## APPENDIX A. STUDY CATCHMENT DETAILS

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>105105A: East Normanby River at Development Road</td>
<td>296.99</td>
<td>145.01</td>
<td>15.77</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>107001B: Endeavour River at Flaggy</td>
<td>341.39</td>
<td>145.07</td>
<td>15.42</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>108002A: Daintree River at Bairds</td>
<td>913.19</td>
<td>145.28</td>
<td>16.18</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>108003A: Bloomfield River at China Camp</td>
<td>263.49</td>
<td>145.29</td>
<td>15.99</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>112002A: Fisher Creek at Nerada</td>
<td>16.08</td>
<td>145.91</td>
<td>17.56</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>112102A: Liverpool Creek at Upper Japoonvale</td>
<td>78.18</td>
<td>145.91</td>
<td>17.72</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>113004A: Cochabue Creek at Powerline</td>
<td>93.72</td>
<td>145.63</td>
<td>17.74</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116006B: Herbert River at Abergowrie</td>
<td>747.50</td>
<td>145.92</td>
<td>18.49</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116010A: Blencoe Creek at Blencoe Falls</td>
<td>227.50</td>
<td>145.54</td>
<td>18.20</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116011A: Millstream River at Ravenshoe</td>
<td>90.94</td>
<td>145.48</td>
<td>17.61</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116012A: Cameron Creek at 8.7KM</td>
<td>352.28</td>
<td>145.34</td>
<td>18.06</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116013A: Millstream river at Archer Creek</td>
<td>320.38</td>
<td>145.34</td>
<td>17.65</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116014A: Wild River at Silver Valley</td>
<td>587.89</td>
<td>145.30</td>
<td>17.63</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>116015A: Blunder Creek at Wooroorra</td>
<td>126.27</td>
<td>145.44</td>
<td>17.74</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>121001A: Don River at Ida Creek</td>
<td>606.21</td>
<td>148.12</td>
<td>20.29</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>122004A: Gregory river at Lower Gregory</td>
<td>46.72</td>
<td>148.55</td>
<td>20.30</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>12003A: Carmila Creek</td>
<td>83.82</td>
<td>149.40</td>
<td>21.92</td>
<td></td>
<td>QL D</td>
<td>Outside of study area - Queensland; too far north</td>
</tr>
<tr>
<td>12002D: Barambah Creek</td>
<td>651.67</td>
<td>152.04</td>
<td>26.30</td>
<td>QL D</td>
<td>MA</td>
<td>Two of four flow double mass curves appear suspect; daily flow series unusual.</td>
</tr>
<tr>
<td>13002A: Barker Creek</td>
<td>311.73</td>
<td>151.81</td>
<td>26.74</td>
<td>QL D</td>
<td>MA</td>
<td>Strong rain contrast but otherwise ok.</td>
</tr>
<tr>
<td>13008A: Boonara Creek</td>
<td>1308.30</td>
<td>151.84</td>
<td>25.90</td>
<td>QL D</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>137101A: Gregory River</td>
<td>454.41</td>
<td>152.24</td>
<td>25.09</td>
<td>QL D</td>
<td>MA</td>
<td>A lot of data missing (8.5%) and FDC different as a result. Three of four flow double mass curves appear suspect, but independent verification from 13721A. Data missing (7.3%) but otherwise ok.</td>
</tr>
<tr>
<td>13721A: Isis River at</td>
<td>445.18</td>
<td>152.37</td>
<td>25.27</td>
<td>QL D</td>
<td>ACC</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>Bruce Highway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>138004B: Munna Creek</td>
<td>1193.75</td>
<td>152.35</td>
<td>25.90</td>
<td>QL D</td>
<td>REJ</td>
<td></td>
</tr>
<tr>
<td>138009A: Tagigan Creek</td>
<td>101.72</td>
<td>152.78</td>
<td>26.08</td>
<td>QL D</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Road</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13810A: Wide Bay Creek</td>
<td>352.30</td>
<td>152.22</td>
<td>26.08</td>
<td>QL D</td>
<td>MA</td>
<td>A lot of data missing (14.3%) but otherwise ok.</td>
</tr>
<tr>
<td>138113A: Kandanga Creek</td>
<td>170.78</td>
<td>152.64</td>
<td>26.39</td>
<td>QL D</td>
<td>MA</td>
<td>A lot of data missing (12.4%) but otherwise ok.</td>
</tr>
<tr>
<td>143009A: Brisbane River</td>
<td>3886.69</td>
<td>152.41</td>
<td>27.00</td>
<td>QL D</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>at Gregors Creek</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>143110A: Bremer River</td>
<td>126.09</td>
<td>152.51</td>
<td>27.83</td>
<td>QL D</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>at Adams Bridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>143303A: Stanley River</td>
<td>104.33</td>
<td>152.84</td>
<td>26.84</td>
<td>QL D</td>
<td>ACC</td>
<td>May be high rainfall contrast - beyond the display gradation I have adopted.</td>
</tr>
<tr>
<td>at Penchester</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX A. STUDY CATCHMENT DETAILS

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>145010A: Running Creek at 5.8KM</td>
<td>132.58</td>
<td>152.89</td>
<td>28.25</td>
<td>QLD</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>Deickmans Bridge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>145011A: Teviot Brook at Croftby</td>
<td>83.01</td>
<td>152.57</td>
<td>28.15</td>
<td>QLD</td>
<td>MA</td>
<td>Low flow behaviour suspect; beware using low flow signatures. Rainfall contrast high.</td>
</tr>
<tr>
<td>145018A: Burnett Creek at Up Stream</td>
<td>81.77</td>
<td>152.61</td>
<td>28.22</td>
<td>QLD</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Maroon Dam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>145101D: Albert River at Lumeah Number 2</td>
<td>167.39</td>
<td>153.05</td>
<td>28.05</td>
<td>QLD</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>145107A: Canungra Creek at Main Rd Bridge</td>
<td>99.84</td>
<td>153.16</td>
<td>28.00</td>
<td>QLD</td>
<td>REJ</td>
<td>High rainfall contrast in area but no rainfall gauges in catchment. Thus, risky.</td>
</tr>
<tr>
<td>146010A: Coomera River at Army Camp</td>
<td>97.30</td>
<td>153.19</td>
<td>28.03</td>
<td>QLD</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>146012A: Currumbin Creek at Nicolls Bridge</td>
<td>30.20</td>
<td>153.42</td>
<td>28.18</td>
<td>QLD</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>146014A: Back Creek at Beechmont</td>
<td>6.84</td>
<td>153.19</td>
<td>28.12</td>
<td>QLD</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>146095A: Tallebudgera Creek at Tallebudgera Ck Rd</td>
<td>56.47</td>
<td>153.40</td>
<td>28.15</td>
<td>QLD</td>
<td>ACC</td>
<td>Four out of four double mass curves have a distinct lean to the right towards the end of the timeseries. Flow gauging error?</td>
</tr>
<tr>
<td>204034: Henry River at Newton Boyd</td>
<td>401.24</td>
<td>152.21</td>
<td>29.76</td>
<td>NSW</td>
<td>REJ</td>
<td></td>
</tr>
<tr>
<td>206014: Wolloomombi River at Cominside</td>
<td>370.92</td>
<td>152.03</td>
<td>30.48</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>206018: Apsley River at Apsley Falls</td>
<td>863.62</td>
<td>151.77</td>
<td>31.05</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>208007: Nowendoc River at Nowendoc</td>
<td>222.74</td>
<td>151.72</td>
<td>31.52</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>208009: Barnard River at Barry</td>
<td>151.74</td>
<td>151.32</td>
<td>31.58</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>210006: Goulburn River at Coggan</td>
<td>3404.50</td>
<td>150.10</td>
<td>32.34</td>
<td>NSW</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>210011: William River at Tillega</td>
<td>196.52</td>
<td>151.69</td>
<td>32.32</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>211008: Jigadoo Creek at Avendale</td>
<td>66.43</td>
<td>151.47</td>
<td>33.07</td>
<td>NSW</td>
<td>ACC</td>
<td>Contrasting topography over catchment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Daily flow looks suspect. Four of four flow double mass curves have a distinct lean to the left towards the end of the timeseries. Flow gauging error?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Majority of catchment has similar rain; lowest 1% near outlet has distinctly lower rainfall.</td>
</tr>
<tr>
<td>212200: Nepean River at Maguires Crossing</td>
<td>70.34</td>
<td>150.53</td>
<td>34.48</td>
<td>NSW</td>
<td>REJ</td>
<td></td>
</tr>
<tr>
<td>212200: Kowmung River at Cedar Ford</td>
<td>716.14</td>
<td>150.24</td>
<td>33.95</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>215002: Shoalhaven River at Warri</td>
<td>1400.59</td>
<td>149.74</td>
<td>35.34</td>
<td>NSW</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>215004: Corang River at Hockeys</td>
<td>163.70</td>
<td>150.03</td>
<td>35.15</td>
<td>NSW</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>216002: Clyde River at Brooman</td>
<td>861.65</td>
<td>150.24</td>
<td>35.47</td>
<td>NSW</td>
<td>MA</td>
<td>A little large, also a lot of days with flow flagged or infilled.</td>
</tr>
<tr>
<td>216004: Currambene Creek at Falls Creek</td>
<td>93.48</td>
<td>150.60</td>
<td>34.97</td>
<td>NSW</td>
<td>ACC</td>
<td>A LOT of flagged data.</td>
</tr>
<tr>
<td>218001: Tuross River at Tuross Vale</td>
<td>90.14</td>
<td>149.51</td>
<td>36.26</td>
<td>NSW</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>219001: Rutherford Creek at Brown Mountain</td>
<td>15.33</td>
<td>149.44</td>
<td>36.59</td>
<td>NSW</td>
<td>REJ</td>
<td>Don't trust the rainfall interpolation; rain looks too low given flow.</td>
</tr>
<tr>
<td>221207: Errinundra River at Errinundra</td>
<td>158.19</td>
<td>148.91</td>
<td>37.45</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>221210: Genoa River at The Gorge</td>
<td>838.55</td>
<td>149.52</td>
<td>37.42</td>
<td>VIC</td>
<td>REJ</td>
<td>A little large, also don't trust the rainfall interpolation.</td>
</tr>
<tr>
<td>222206: Buchan River at Buchan</td>
<td>850.78</td>
<td>148.17</td>
<td>37.49</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>22213: Suggan Buggan River at Suggan Buggan</td>
<td>364.47</td>
<td>148.33</td>
<td>36.95</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast high, this is why it appears at the top of the Budyko curve.</td>
</tr>
<tr>
<td>223202: Tambo River at Swifts Creek</td>
<td>897.17</td>
<td>147.73</td>
<td>37.27</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>224206: Wonnangatta River at Crooked River</td>
<td>1106.33</td>
<td>147.09</td>
<td>37.41</td>
<td>VIC</td>
<td>MA</td>
<td>Rainfall contrast is strong; AE close to PE. Only accepted because so many from this area are rejected - include one for representativeness.</td>
</tr>
<tr>
<td>224213A: Dargo River at Lower Dargo Rd</td>
<td>668.77</td>
<td>147.27</td>
<td>37.49</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>224214A: Wentworth River at Tabberabbera</td>
<td>441.65</td>
<td>147.39</td>
<td>37.49</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>225020A: South Cascade Creek at Thompson Valley Road</td>
<td>11.53</td>
<td>146.35</td>
<td>37.83</td>
<td>VIC</td>
<td>REJ</td>
<td>Many things uncertain. Extremely small catchment; boundary / size uncertain. No rain gauges nearby; flow looks too high or rainfall too low.</td>
</tr>
<tr>
<td>225110A: Jordan River at D/S Johnsons Creek</td>
<td>132.17</td>
<td>146.32</td>
<td>37.68</td>
<td>VIC</td>
<td>REJ</td>
<td>AE = PE. PE too low?</td>
</tr>
<tr>
<td>225219: Macalister River at Glencairn</td>
<td>573.61</td>
<td>146.57</td>
<td>37.52</td>
<td>VIC</td>
<td>MA</td>
<td>Strong rain contrast but otherwise ok.</td>
</tr>
<tr>
<td>226220: Loch River at Noojee</td>
<td>104.31</td>
<td>146.01</td>
<td>37.87</td>
<td>VIC</td>
<td>REJ</td>
<td>AE &gt; PE. PE too low?</td>
</tr>
<tr>
<td>226222: Latrobe River at Near Noojee</td>
<td>64.71</td>
<td>145.89</td>
<td>37.88</td>
<td>VIC</td>
<td>MA</td>
<td>AE nearly equal to PE. PE too low?</td>
</tr>
<tr>
<td>226407: Morwell River at Boolarra</td>
<td>115.12</td>
<td>146.30</td>
<td>38.41</td>
<td>VIC</td>
<td>REJ</td>
<td>Suspect short term kink in flow double mass curves, all four. Flow gauge error?</td>
</tr>
<tr>
<td>227225A: Tarra River at Fischers</td>
<td>20.27</td>
<td>146.56</td>
<td>38.47</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>----</td>
<td>------</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>East Branch at Dumbalk North</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>227227: Willkur Creek at Leongatha</td>
<td>105.19</td>
<td>145.96</td>
<td>38.39</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>229650A: Aldermans Creek at RD 32</td>
<td>25.49</td>
<td>145.94</td>
<td>37.72</td>
<td>VIC</td>
<td>REJ</td>
<td>AE &gt; PE. PE too low?</td>
</tr>
<tr>
<td>229661A: Walshes Creek at RD 1</td>
<td>54.47</td>
<td>145.93</td>
<td>37.63</td>
<td>VIC</td>
<td>REJ</td>
<td>AE &gt; PE. PE too low?</td>
</tr>
<tr>
<td>230210: Saltwater Creek at Bullengarook</td>
<td>38.94</td>
<td>144.52</td>
<td>37.47</td>
<td>VIC</td>
<td>REJ</td>
<td>All four flow double mass curves show a similar inflection. Flow gauge error?</td>
</tr>
<tr>
<td>231213: Lederderg River at Sardine Creek</td>
<td>133.36</td>
<td>144.36</td>
<td>37.50</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>O’Brien Crossing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>233205: Arkins Creek</td>
<td>4.42</td>
<td>143.44</td>
<td>38.64</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>236213: Mount Emu Creek at Mena Park</td>
<td>464.57</td>
<td>143.47</td>
<td>37.53</td>
<td>VIC</td>
<td>REJ</td>
<td>Suspect that isolated pockets of high rainfall are producing much of the flow. Don’t trust the catchment delineation; checked against topo map, disappearing streams?</td>
</tr>
<tr>
<td>238208: Jimmy Creek at Jimmy Creek</td>
<td>22.51</td>
<td>142.51</td>
<td>37.37</td>
<td>VIC</td>
<td>REJ</td>
<td></td>
</tr>
<tr>
<td>302214: Anson’s River at Downstream of Big Boggy Creek</td>
<td>229.31</td>
<td>148.22</td>
<td>41.05</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>304497: Nive River at Gowan Brae</td>
<td>177.69</td>
<td>146.42</td>
<td>42.03</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>304499: Tyenna River at Newbury</td>
<td>206.11</td>
<td>146.71</td>
<td>42.71</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>305202: Smug Rivulet at U/S Smug Tiers Rd Bridge</td>
<td>17.41</td>
<td>147.24</td>
<td>43.07</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>307473: Davey River at D/S Crossing Rv</td>
<td>307.54</td>
<td>145.95</td>
<td>43.14</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>308145: Franklin River at Mt Finham Track</td>
<td>772.14</td>
<td>145.77</td>
<td>42.24</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>308799: Collingwood river at B/L Alma</td>
<td>268.32</td>
<td>145.93</td>
<td>42.16</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>312061: Heliker River at Guilford Junction</td>
<td>99.41</td>
<td>145.67</td>
<td>41.42</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>314207: Leven River at at Bannons Bridge</td>
<td>500.59</td>
<td>146.09</td>
<td>41.25</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>314213: Black River at South Forest</td>
<td>319.31</td>
<td>145.30</td>
<td>40.87</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>315450: Forth River at U/S Lemonyhme</td>
<td>309.91</td>
<td>146.13</td>
<td>41.61</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>318076: North Esk River at Ballroom</td>
<td>374.55</td>
<td>147.38</td>
<td>41.49</td>
<td>TAS</td>
<td>REJ</td>
<td>Outside of study area - Tasmania</td>
</tr>
<tr>
<td>401009: Maragle Creek at Maragle</td>
<td>215.93</td>
<td>148.10</td>
<td>35.93</td>
<td>NSW</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>401012: Murray River at Biggarra</td>
<td>1262.92</td>
<td>148.05</td>
<td>36.32</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>401015: Bowna Creek at Yambala</td>
<td>274.16</td>
<td>146.98</td>
<td>35.92</td>
<td>NSW</td>
<td>MA</td>
<td>A LOT of flagged data.</td>
</tr>
<tr>
<td>401203: Mitta Mitta River at Hinnomunjie</td>
<td>1526.80</td>
<td>147.61</td>
<td>36.95</td>
<td>VIC</td>
<td>REJ</td>
<td>A bit big; rain contrast; AE = PE (PE to low?)</td>
</tr>
<tr>
<td>401208: Cudgewa Creek at Berringama</td>
<td>357.02</td>
<td>147.68</td>
<td>36.21</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>401210: Snowy Creek at Below Granite Flat</td>
<td>413.75</td>
<td>147.41</td>
<td>36.57</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>401212: Nariel Creek at Upper Nariel</td>
<td>255.14</td>
<td>147.83</td>
<td>36.45</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>401216: Big River at Jokers Creek</td>
<td>357.87</td>
<td>147.47</td>
<td>36.93</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>401217: Gibbo River at Gibbo Park</td>
<td>389.85</td>
<td>147.71</td>
<td>36.76</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>402204: Yakandandah Creek at Osbornes Flat</td>
<td>280.97</td>
<td>146.91</td>
<td>36.30</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>402206: Running Creek at Running Creek Flat</td>
<td>127.56</td>
<td>147.04</td>
<td>36.54</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>402213: Kinchington Creek at Osbornes Flat</td>
<td>120.34</td>
<td>146.89</td>
<td>36.32</td>
<td>VIC</td>
<td>REJ</td>
<td>We'll use the gauge just downstream instead (4224)</td>
</tr>
<tr>
<td>402217: Flaggy Creek at Myrtleford Road Bridge</td>
<td>22.55</td>
<td>146.88</td>
<td>36.39</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>403209A: Reedy Creek at Wangaratta North</td>
<td>5521.88</td>
<td>146.34</td>
<td>36.33</td>
<td>VIC</td>
<td>REJ</td>
<td>Appears to be an error in either gauge location or catchment declination.</td>
</tr>
<tr>
<td>403213A: Fifteen Mile Creek at Greta South</td>
<td>227.64</td>
<td>146.24</td>
<td>36.62</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast but otherwise ok.</td>
</tr>
<tr>
<td>403214: Happy Valley Creek at Rosewhite</td>
<td>137.81</td>
<td>146.82</td>
<td>36.58</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>403217: Rose River at Matong North</td>
<td>178.16</td>
<td>146.58</td>
<td>36.82</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>403221: Reedy Creek at Woolshed</td>
<td>211.74</td>
<td>146.60</td>
<td>36.31</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>403222: Buffalo River at Abbeyard</td>
<td>415.71</td>
<td>146.70</td>
<td>36.91</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>403226: Boggy Creek at Angleside</td>
<td>108.10</td>
<td>146.36</td>
<td>36.61</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>403232: Morses Creek at Wandiligong</td>
<td>125.53</td>
<td>146.98</td>
<td>36.75</td>
<td>VIC</td>
<td>REJ</td>
<td>AE nearly equal to PE. PE too low?</td>
</tr>
<tr>
<td>404207: Holland Creek at Kelleera</td>
<td>455.88</td>
<td>146.06</td>
<td>36.61</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>405205: Murrindindi River at Murrindindi above Colwells</td>
<td>107.96</td>
<td>145.56</td>
<td>37.41</td>
<td>VIC</td>
<td>MA</td>
<td>AE nearly equal to PE. PE too low?</td>
</tr>
</tbody>
</table>
### APPENDIX A: STUDY CATCHMENT DETAILS

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>405209: Acheron River at Taggerty</td>
<td>627.11</td>
<td>145.71</td>
<td>37.32</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast. AE nearly equal to PE. PE too low?</td>
</tr>
<tr>
<td>405215: Howqua River at Glen Esk</td>
<td>370.06</td>
<td>146.21</td>
<td>37.23</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>405217: Yea River at Devlins Bridge</td>
<td>368.74</td>
<td>145.47</td>
<td>37.38</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast, but each trib has same profile.</td>
</tr>
<tr>
<td>405218: Jamieson River at Gerring Bridge</td>
<td>367.43</td>
<td>146.19</td>
<td>37.29</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>405219: Goulburn River at Dohertys</td>
<td>700.84</td>
<td>146.13</td>
<td>37.33</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>405226: Pranip Creek at Moorlim</td>
<td>788.68</td>
<td>145.31</td>
<td>36.62</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>405227: Big River at Jamieson</td>
<td>629.70</td>
<td>146.06</td>
<td>37.37</td>
<td>VIC</td>
<td>MA</td>
<td>AE nearly equal to PE. PE too low?</td>
</tr>
<tr>
<td>405238: Mollison Creek at Pyalong</td>
<td>164.39</td>
<td>144.86</td>
<td>37.12</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>405245: Ford Creek at Mansfield</td>
<td>117.43</td>
<td>146.05</td>
<td>37.04</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>405248: Major Creek at Graytown</td>
<td>291.78</td>
<td>144.91</td>
<td>36.85</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>405251: Brankeet Creek at Anconia</td>
<td>118.70</td>
<td>145.78</td>
<td>36.97</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>405263: Goulburn River at U/S of Snake Creek Junction</td>
<td>328.09</td>
<td>146.25</td>
<td>37.46</td>
<td>VIC</td>
<td>REJ</td>
<td>We'll use the gauge downstream instead (45219)</td>
</tr>
<tr>
<td>405264: Big River at D/S of Frenchman Creek Junction</td>
<td>337.07</td>
<td>146.08</td>
<td>37.52</td>
<td>VIC</td>
<td>REJ</td>
<td>We'll use the gauge downstream instead (45227)</td>
</tr>
<tr>
<td>405274: Home Creek at Yarck</td>
<td>181.65</td>
<td>145.61</td>
<td>37.11</td>
<td>VIC</td>
<td>MA</td>
<td>Flow double mass curves look slightly suspect. Otherwise ok.</td>
</tr>
<tr>
<td>406208: Campaspe River at Ashborne</td>
<td>37.83</td>
<td>144.45</td>
<td>37.39</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>406213: Campaspe River at Redesdale</td>
<td>638.54</td>
<td>144.54</td>
<td>37.02</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast, but included as an important example catchment with respect to the Millennium Drought.</td>
</tr>
<tr>
<td>406214: Axe Creek at Longlea</td>
<td>230.20</td>
<td>144.43</td>
<td>36.77</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast but otherwise ok</td>
</tr>
<tr>
<td>406224: Mount Pleasant Creek at Runnymede</td>
<td>242.60</td>
<td>144.64</td>
<td>36.55</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>407214: Creswick Creek at Clunes</td>
<td>305.79</td>
<td>143.79</td>
<td>37.30</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>407215: Loddon River at Newstead</td>
<td>1029.95</td>
<td>144.06</td>
<td>37.11</td>
<td>VIC</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>407220: Bet Bet Creek at Norwood</td>
<td>351.68</td>
<td>143.64</td>
<td>36.99</td>
<td>VIC</td>
<td>MA</td>
<td>Flow double mass curves look slightly suspect. Otherwise ok.</td>
</tr>
<tr>
<td>407230: Joycees Creek at Strathlea</td>
<td>150.81</td>
<td>143.96</td>
<td>37.16</td>
<td>VIC</td>
<td>MA</td>
<td>High rainfall contrast but otherwise ok</td>
</tr>
<tr>
<td>407253: Piccaninny Creek at Minto</td>
<td>680.65</td>
<td>144.47</td>
<td>36.45</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>408200: Avoca River at Coonooer</td>
<td>2689.47</td>
<td>143.30</td>
<td>36.44</td>
<td>VIC</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>408202: Avoca River at Amphithetre</td>
<td>76.60</td>
<td>143.41</td>
<td>37.18</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>410057: Goobarragandra River at Lacmalac</td>
<td>665.78</td>
<td>148.35</td>
<td>35.33</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>410061: Adelong Creek at Batlow Road</td>
<td>146.88</td>
<td>148.07</td>
<td>35.33</td>
<td>NSW</td>
<td>ACC</td>
<td>A LOT of flagged data!</td>
</tr>
<tr>
<td>410705: Molonglo River at Burbong</td>
<td>498.85</td>
<td>149.32</td>
<td>35.34</td>
<td>NSW</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>410730: Cotter River at Gingera</td>
<td>129.31</td>
<td>148.82</td>
<td>35.59</td>
<td>ACT</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>410731: Gudgenby River at Tennent</td>
<td>671.10</td>
<td>149.07</td>
<td>35.57</td>
<td>ACT</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>410734: Queanbeyan River at Tinderry</td>
<td>565.18</td>
<td>149.35</td>
<td>35.61</td>
<td>NSW</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------</td>
<td>-------</td>
<td>-------</td>
<td>----</td>
<td>-----</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>410761: Murumbidgee River at Lobb Holes</td>
<td>5190.18</td>
<td>149.10</td>
<td>35.54</td>
<td>ACT</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>412028: Abercrombie River at Abercrombie</td>
<td>2651.52</td>
<td>149.33</td>
<td>33.95</td>
<td>NSW</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>412050: Crookwell River at Narraw North</td>
<td>760.55</td>
<td>149.17</td>
<td>34.31</td>
<td>NSW</td>
<td>REJ</td>
<td>Suspect flow double mass curves, all four. Flow gauge error?</td>
</tr>
<tr>
<td>412066: Abercrombie River at Hadley No. 2</td>
<td>1641.50</td>
<td>149.60</td>
<td>34.11</td>
<td>NSW</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>415207: Wimmera River at Eversley</td>
<td>306.05</td>
<td>143.18</td>
<td>37.19</td>
<td>VIC</td>
<td>REJ</td>
<td>Suspect flow double mass curves, three out of four. Flow gauge error? Also high rainfall contrast.</td>
</tr>
<tr>
<td>415226: Richardson River at Carrs Plains</td>
<td>129.78</td>
<td>142.79</td>
<td>36.74</td>
<td>VIC</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>415237: Concongella Creek at Stawell</td>
<td>239.91</td>
<td>142.82</td>
<td>37.03</td>
<td>VIC</td>
<td>MA</td>
<td>Flow double mass curves look slightly suspect. Otherwise ok.</td>
</tr>
<tr>
<td>416003: Tenterfield Creek at Clifton</td>
<td>556.71</td>
<td>151.73</td>
<td>29.03</td>
<td>NSW</td>
<td>MA</td>
<td>Flow appears quite low, plus high rainfall contrast. Otherwise ok.</td>
</tr>
<tr>
<td>416008: Beadt River at Haystack</td>
<td>900.73</td>
<td>151.38</td>
<td>29.22</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>418005: Copes Creek at Kimberley</td>
<td>237.95</td>
<td>151.11</td>
<td>29.92</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>418014: Gwydir River at Yarrowyck</td>
<td>824.52</td>
<td>151.36</td>
<td>30.47</td>
<td>NSW</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>419005: Namoi River at North Cuerindi</td>
<td>2544.37</td>
<td>150.78</td>
<td>30.68</td>
<td>NSW</td>
<td>REJ</td>
<td>Rain contrast too large due to large catchment area</td>
</tr>
<tr>
<td>422202B: Dogwood Creek at Gilweir</td>
<td>2984.31</td>
<td>150.18</td>
<td>26.71</td>
<td>QLD</td>
<td>REJ</td>
<td>Suspect flow double mass curves, all four. Flow gauge error?</td>
</tr>
<tr>
<td>422306A: Swan Creek at Swanfels</td>
<td>82.10</td>
<td>152.28</td>
<td>28.16</td>
<td>QLD</td>
<td>MA</td>
<td>Flow double mass curves look slightly suspect. High rainfall contrast. Low flow looks suspect - beware using low flow signatures.</td>
</tr>
<tr>
<td>422313B: Emu Creek at Emu Vale</td>
<td>153.79</td>
<td>152.23</td>
<td>28.23</td>
<td>QLD</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>422319B: Dalrymple Creek at Allora</td>
<td>246.33</td>
<td>152.01</td>
<td>28.04</td>
<td>QLD</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>4223121B: Spring Creek at Killarney</td>
<td>33.41</td>
<td>152.33</td>
<td>28.35</td>
<td>QLD</td>
<td>ACC</td>
<td>Flow double mass curves look slightly suspect, but otherwise ok.</td>
</tr>
<tr>
<td>422334A: Kings Creek at Aides Bridge</td>
<td>514.85</td>
<td>151.86</td>
<td>27.93</td>
<td>QLD</td>
<td>MA</td>
<td></td>
</tr>
<tr>
<td>422391A: Condamine River at Elbow Valley</td>
<td>328.86</td>
<td>152.14</td>
<td>28.37</td>
<td>QLD</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>421002: Paroo River at Willarra Crossing</td>
<td>35100.63144.46</td>
<td>29.24</td>
<td>NSW</td>
<td>REJ</td>
<td></td>
<td>Outside of study area - inland QLD</td>
</tr>
<tr>
<td>421201A: Paroo River at Caiwarro</td>
<td>22715.32144.79</td>
<td>28.69</td>
<td>QLD</td>
<td>REJ</td>
<td></td>
<td>Outside of study area - inland QLD</td>
</tr>
<tr>
<td>604053: Kent River at Styx Junction</td>
<td>1849.35</td>
<td>117.09</td>
<td>34.89</td>
<td>WA</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>606001: Deep River at Ted's Pool</td>
<td>475.04</td>
<td>116.62</td>
<td>34.77</td>
<td>WA</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>606002: Weld River at Wattle Block</td>
<td>60.58</td>
<td>116.52</td>
<td>34.69</td>
<td>WA</td>
<td>REJ</td>
<td>Rainfall contrast too high, Anomaly - runoff unusually low (rain = 13mm/yr; flow = 36mm/yr). Measurement error? Geology?</td>
</tr>
<tr>
<td>606185: Shannon River at Dog Pool</td>
<td>433.18</td>
<td>116.38</td>
<td>34.76</td>
<td>WA</td>
<td>MA</td>
<td>High rainfall contrast but otherwise ok</td>
</tr>
<tr>
<td>607150: Dombaleup Brook at Malimup Track</td>
<td>125.09</td>
<td>115.97</td>
<td>34.58</td>
<td>WA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>608002: Carey Brook at Staircase Road</td>
<td>44.52</td>
<td>115.84</td>
<td>34.39</td>
<td>WA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>610008: Margaret River North at Whicher Range</td>
<td>15.72</td>
<td>115.44</td>
<td>33.81</td>
<td>WA</td>
<td>REJ</td>
<td>Flow double mass curves very suspect.</td>
</tr>
<tr>
<td>613002: Harvy River at Dingo Road</td>
<td>147.47</td>
<td>116.04</td>
<td>33.09</td>
<td>WA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>613146: Clarke Brook at Hillview Farm</td>
<td>16.94</td>
<td>115.92</td>
<td>33.00</td>
<td>WA</td>
<td>ACC</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX A. STUDY CATCHMENT DETAILS

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>614044: Yarragil Brook at Yarragil Formation</td>
<td>73.33</td>
<td>116.15</td>
<td>32.81</td>
<td>WA</td>
<td>REJ</td>
<td>Anomaly - runoff unusually low (rain = 943mm/yr; flow = 38mm/yr). Measurement error? Geology?</td>
</tr>
<tr>
<td>616002: Darkin River at Pine Plantation</td>
<td>671.24</td>
<td>116.29</td>
<td>32.07</td>
<td>WA</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>616013: Helena River at Ngangaguringing</td>
<td>328.76</td>
<td>116.40</td>
<td>31.94</td>
<td>WA</td>
<td>REJ</td>
<td>Practically no flow, runoff ratio .01.</td>
</tr>
<tr>
<td>616055: Canning River at Glen Eagle</td>
<td>520.87</td>
<td>116.17</td>
<td>32.23</td>
<td>WA</td>
<td>REJ</td>
<td>Rainfall contrast too high</td>
</tr>
<tr>
<td>803003: Fletcher River at Dromedary</td>
<td>65.31</td>
<td>124.99</td>
<td>17.13</td>
<td>WA</td>
<td>REJ</td>
<td>Outside of study area - Northern WA</td>
</tr>
<tr>
<td>804001: Isdell River at Dales Yard</td>
<td>1841.63</td>
<td>125.43</td>
<td>17.01</td>
<td>WA</td>
<td>REJ</td>
<td>Outside of study area - Northern WA</td>
</tr>
<tr>
<td>809310: Ord River at Bedford Downs</td>
<td>550.06</td>
<td>127.60</td>
<td>17.43</td>
<td>WA</td>
<td>REJ</td>
<td>Outside of study area - Northern WA</td>
</tr>
<tr>
<td>912101A: Gregory River at Gregory Downs</td>
<td>12556.17</td>
<td>139.25</td>
<td>18.64</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>912105A: Gregory River at Riversleigh No. 2</td>
<td>11375.16</td>
<td>138.80</td>
<td>18.97</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>915011A: Porcupine Creek at Mt Emu Plains</td>
<td>566.87</td>
<td>144.52</td>
<td>18.18</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>917107A: Elizabeth Creek at Mount Surprise</td>
<td>459.17</td>
<td>144.31</td>
<td>18.13</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>919003A: Mitchell River at O.K. Bridge</td>
<td>7743.10</td>
<td>144.29</td>
<td>16.47</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>919201A: Palmer River at Goldfields</td>
<td>333.69</td>
<td>144.78</td>
<td>16.11</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>919309A: Walsh River at Trimbles Crossing</td>
<td>8684.39</td>
<td>143.78</td>
<td>16.54</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>922101B: Coen river at Racecourse</td>
<td>171.78</td>
<td>143.18</td>
<td>13.96</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>923001A: Wenlock river at Moreton</td>
<td>3313.49</td>
<td>142.64</td>
<td>12.45</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>923002A: Dulhunty River at Dougs Pad</td>
<td>331.63</td>
<td>142.42</td>
<td>11.83</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>A0020101: Diamantina River at Birdsville</td>
<td>121316.2030.37</td>
<td>25.91</td>
<td>QLD</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory / QLD</td>
<td></td>
</tr>
<tr>
<td>A0030501: Cooper Creek at Cullyamurra</td>
<td>237845.3140.84</td>
<td>27.70</td>
<td>SA</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory / QLD</td>
<td></td>
</tr>
<tr>
<td>A2990519: Mosquito Creek at Struan</td>
<td>1185.23</td>
<td>140.78</td>
<td>37.09</td>
<td>SA</td>
<td>REJ</td>
<td>Uncertainty as to catchment boundary</td>
</tr>
<tr>
<td>A2990523: Stony Creek at Woakwine Range</td>
<td>8.39</td>
<td>140.36</td>
<td>37.70</td>
<td>SA</td>
<td>REJ</td>
<td>Uncertainty as to catchment boundary</td>
</tr>
<tr>
<td>A2990531: Morambro Creek at Bordertown-Naracoorte Road Bridge</td>
<td>405.47</td>
<td>140.66</td>
<td>36.69</td>
<td>SA</td>
<td>REJ</td>
<td>Anomaly - runoff unusually low (rain = 521mm/yr; flow = 8mm/yr). Measurement error? Geology?</td>
</tr>
<tr>
<td>A3030502: Scott Creek at Scott Bottom</td>
<td>26.84</td>
<td>138.67</td>
<td>35.10</td>
<td>SA</td>
<td>MA</td>
<td>High rainfall contrast but otherwise ok</td>
</tr>
<tr>
<td>A5040517: First Creek at Waterfall Gully</td>
<td>5.12</td>
<td>138.68</td>
<td>34.97</td>
<td>SA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>A5040523: Sixth Creek at Castambul</td>
<td>43.50</td>
<td>138.75</td>
<td>34.87</td>
<td>SA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>A5030517: North Para River at Penrice</td>
<td>136.42</td>
<td>139.06</td>
<td>34.46</td>
<td>SA</td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td>A5130501: Rocky River upstream Gorge Falls</td>
<td>190.73</td>
<td>136.70</td>
<td>35.95</td>
<td>SA</td>
<td>ACC</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX A. STUDY CATCHMENT DETAILS

<table>
<thead>
<tr>
<th>Name</th>
<th>CA</th>
<th>LON</th>
<th>LAT</th>
<th>ST</th>
<th>DEC</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>G0010005:</td>
<td>Ranken River at Soudan Homestead</td>
<td>66°36.09' 137°02.00' 20°05'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G0050115:</td>
<td>Hugh River at South Road Crossing</td>
<td>33°53.22' 133°43.35' 24°35'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G0060005:</td>
<td>Trephina Creek at Trephina Gorge</td>
<td>43°70.00' 134°38.00' 23°33'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8110004:</td>
<td>East Baines River at VIC Highway</td>
<td>24°51.48' 130°03.00' 15°77'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8110016:</td>
<td>Upper VIC River at Wave Hill Police Station</td>
<td>46°08.55' 130°84.00' 17°45'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8140001:</td>
<td>Katherine River at Railway Bridge</td>
<td>82°28.34' 132°26.00' 14°46'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8140010:</td>
<td>Daly River at Mount Nancar</td>
<td>47°46.21' 130°74.74' 13°83'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8140161:</td>
<td>Green Ant Creek at Tipperary</td>
<td>41°34.00' 131°10.00' 13°74'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8150018:</td>
<td>Elizabeth River at Stuart Highway</td>
<td>96°23.00' 131°07.00' 12°61'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8170002:</td>
<td>Adelaide River at Railway Bridge</td>
<td>64°56.95' 131°11.00' 13°24'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8190001:</td>
<td>West Alligator River at Upstream Arnhem Highway</td>
<td>25°55.18' 132°17.00' 12°79'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>G8200015:</td>
<td>South Alligator River at El Sherana</td>
<td>12°23.88' 132°52.00' 13°52'</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>CA</td>
<td>LON</td>
<td>LAT</td>
<td>ST</td>
<td>DEC</td>
<td>Notes</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>----</td>
<td>-----</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>G8210010: East Alligator River at 12 Deg 43 Minutes South</td>
<td>2395.06</td>
<td>133.33</td>
<td>12.72</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>G9030124: Daly Waters Creek at Daly Waters</td>
<td>197.38</td>
<td>133.37</td>
<td>16.26</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>G9030250: Ropa River at Red Rock</td>
<td>43719.13</td>
<td>134.42</td>
<td>14.70</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
<tr>
<td>G9070142: McArthur River Upstream of Bailey's Grave</td>
<td>4038.07</td>
<td>135.76</td>
<td>16.78</td>
<td>NT</td>
<td>REJ</td>
<td>Outside of study area - Northern Territory</td>
</tr>
</tbody>
</table>
Appendix B

Supplementary material for
Chapter 4

B.1 Introduction

This appendix contains the following:

1. Text discussing categorization of model failure based on the shape of the Pareto Curve (Section B.2, supported by Figure B.1). This text expands upon the summary provided in Section 4.4.4.

2. Text regarding bushfires (Section B.3) outlining efforts to investigate whether the hydrological effects of bushfires might be associated with catchments where no model structure could meet a given modelling standard. Referenced in Section 4.4.4.

3. Selected figures and tables:

   (a) Pareto results for every case study examined in Chapter 4 (Figures B.2 and B.3). Specifically, we provide a plot for each catchment showing the Pareto curve obtained for each of the five model structures, for a total of 430 curves.
(b) Comparison table for Kling Gupta Efficiency versus Nash Sutcliffe Efficiency (Figure B.4) highlighting problems noted by Gupta et al. (2009).

(c) Plotted comparison of calibration results between AMALGAM and the single-objective optimizer CMA-ES (Figure B.5). This comparison was done in ten catchments, for five model structures, for each of the two objectives.

(d) Four additional case studies pursuant to Section 4.4.2 (Figure B.6). Section 4.4.2 showed an example (Figure 4.8) where the endpoints of two Pareto curves (those for GR4J and GR4JMOD) were similar despite divergence mid-curve. As Section 4.4.2 explains, “...use of a single-objective DSST... would lead to the erroneous conclusion that the alterations to GR4JMOD by Hughes et al. (2013) made negligible difference to the model’s capabilities. In contrast, Figure 4.8 shows that this is not the case by the divergence of the purple GR4JMOD curve from the orange GR4J curve.” To demonstrate that Figure 4.8 is not an isolated case, in Figure B.6 we provide examples from four other catchments.

(e) Catchment-by-catchment information for the 12 catchments in which no model structure was able to meet modelling standard 1 (Table B.1). We provide additional information for the interested reader who wishes to know the names and physical properties of the catchments that were most difficult to model.

(f) Catchment location map (Figure B.7) showing location of catchments where no model structure could meet a given modelling standard, referenced in Section 4.4.2.

(g) Diagrams of simulation bias for selected parameter sets (Figures B.8 to B.11), following a format suggested by Coron et al. (2014). These plots are the same as Figure 4.12, but are shown for more case studies.
B.2 Categorisation of model failure based on shape of the Pareto Curve

In this section we examine those catchments where the model structures failed to meet Standard 1 and/or Standard 2. We analyse the Pareto Curves and consider what the form of these curves may indicate about the type of model failure.

Figure B.1 divides the instances where model structures failed to meet the standards (ie. Case C in Figure B.1) into sub-categories according the type of failure, which we deduce from the form of their Pareto Curves. Having failed to meet the standard, every instance listed in these tables is one where no single parameter set can simulate flows satisfactorily in both wet and dry periods. The endpoints of the curves indicate whether the model structure is able to meet the standard in a given objective when optimized to it in isolation. If it can meet one or both standards in isolation, we classify this as a different type of failure compared to instances where the standard can be met in neither objective.

Figure B.1: Details of model failure for those catchments where a given model structure failed.
The results are relatively well spread between the failure types, making generalisation difficult (Figure B.1). GR4J and GR4JMOD were exceptionally good at meeting the lower standard (KGE = 0.7) in a given objective provided that they were calibrated to it in isolation; hence zero catchments in Failure Type 1 (see also Figure B.1). IHACRES had a much lower failure rate (hence lower totals) but among failure instances, Failure Type 1 was relatively common; that is, IHACRES failures tended to belong to the “appears to be deficient in this catchment, regardless of climate” category, particularly for the more stringent of the two standards (Figure B.1). This category was also common for the model structure SIMHYD. It was more common for model structures to fail to meet the KGE=0.7 standard in the dry period (Failure types 3 and 1) than in the wet (Failure types 2 and 1). This distinction was not evident for KGE = 0.8. In general, the lack of a dominating category, particularly for the higher standard, speaks against the generalisation that model structures are generally poor at simulating dry conditions.

B.3 Discussion of bushfires

We considered the possible effect of bushfires on the hydrology of the test catchments. As with farm dams, data on historic bushfire severity was only available for the State of Victoria (Department of Environment, Land, Water and Planning, 2015c). There have been a number of significant bushfires in Victoria over the last 15 years, notably in 2003, 2007 and 2009. Also relevant to this study was a significant fire in 1983. Bushfires in Australia have been shown to cause changes in catchment hydrology, including a decrease in plant water use immediately following a fire, followed by a period of higher water use as saplings recolonise (eg. Kuczera, 1987; Cornish and Vertessy, 2001). These changes may cause temporary degradation of performance of a calibrated model. In the current study, bushfires could confound the results if the 7-year dry period immediately followed a bushfire. If this was the case, bushfires could be a reason for model failure in the dry period. Given that the modelling period adopted in this study runs to 2010, the 2007 and 2009 fires occurred too late in the
Table B.1: Characteristics for each of the 12 study catchments in group FF; that is, for which no model structure was able to meet modelling standard 1 (KGE_{non-dry} = 0.7; KGE_{dry} = 0.7). Farm dam development data are only available for catchments in the state of Victoria.

<table>
<thead>
<tr>
<th>Catchment</th>
<th>Area (km²)</th>
<th>Mean precip. (mm/yr)</th>
<th>Dry period flow ratio</th>
<th>Average slope (%)</th>
<th>Forest Cover (%)</th>
<th>Farm Dam Development (ML/km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>208009 - Barnard River at Barry</td>
<td>152</td>
<td>958</td>
<td>0.766</td>
<td>23.1</td>
<td>98.4</td>
<td>n/a</td>
</tr>
<tr>
<td>216002 - Clyde River at Brooman</td>
<td>862</td>
<td>1066</td>
<td>0.521</td>
<td>22.7</td>
<td>97.4</td>
<td>n/a</td>
</tr>
<tr>
<td>405248 - Major Creek at Graytown</td>
<td>292</td>
<td>608</td>
<td>0.061</td>
<td>4.3</td>
<td>72.3</td>
<td>4.13</td>
</tr>
<tr>
<td>406214 - Axe Creek at Longlea</td>
<td>236</td>
<td>600</td>
<td>0.068</td>
<td>5.0</td>
<td>71.4</td>
<td>32.06</td>
</tr>
<tr>
<td>406224 - Mount Pleasant Ck at Runnymede</td>
<td>243</td>
<td>518</td>
<td>0.053</td>
<td>2.4</td>
<td>21.4</td>
<td>6.93</td>
</tr>
<tr>
<td>407253 - Piccaninny Creek at Minto</td>
<td>681</td>
<td>491</td>
<td>0.441</td>
<td>1.5</td>
<td>59.3</td>
<td>22.54</td>
</tr>
<tr>
<td>408202 - Avoca River at Amphitheatre</td>
<td>77</td>
<td>616</td>
<td>0.125</td>
<td>13.3</td>
<td>49.5</td>
<td>5.35</td>
</tr>
<tr>
<td>415226 - Richardson River at Carrs Plains</td>
<td>130</td>
<td>504</td>
<td>0.002</td>
<td>1.3</td>
<td>17.9</td>
<td>2.58</td>
</tr>
<tr>
<td>136202D - Barambah Creek at Litzows</td>
<td>652</td>
<td>876</td>
<td>0.158</td>
<td>9.1</td>
<td>96.1</td>
<td>n/a</td>
</tr>
<tr>
<td>137101A - Gregory River at Burrum Hwy</td>
<td>454</td>
<td>918</td>
<td>0.154</td>
<td>4.1</td>
<td>98.3</td>
<td>n/a</td>
</tr>
<tr>
<td>138010A - Wide Bay Creek at Kilkivan</td>
<td>352</td>
<td>819</td>
<td>0.056</td>
<td>10.4</td>
<td>97.7</td>
<td>n/a</td>
</tr>
<tr>
<td>422306A - Swan Creek at Swanfels</td>
<td>82</td>
<td>1012</td>
<td>0.345</td>
<td>24.2</td>
<td>96.4</td>
<td>n/a</td>
</tr>
</tbody>
</table>

record (ie. are too recent) for this confounding to occur. To examine the 2003 and 1983 fires, we looked for catchments where the 7-year period was immediately after the fire. Although many catchments had dry periods starting around 2003 (Figure 4.4), only three of these catchments were within the extent of the 2003 bushfires (401217, 403222 and 402204) and none of these three failed Modelling Standards 1 or 2. Very few catchments had dry periods starting around 1983 (Figure 4.4; only one of these was in Victoria (407214) and this catchment was not burned in the 1983 fires. Based on the above discussion, it is concluded that there is no evidence that poor model performance observed in Groups PF and FF is the result of bushfires.

B.4 Selected figures and tables
Figure B.2: Pareto curves, for each of the five model structures, in 48 of the 86 study catchments. The y axis is KGE over non-dry period (axis limits: 0,1) and the x axis is KGE over dry period (axis limits: 0,1). For plot legend, refer next figure.
Figure B.3: Pareto curves, for each of the five model structures, in 38 of the 86 study catchments. The y axis is KGE over non-dry period (axis limits: 0,1) and the x axis is KGE over dry period (axis limits: 0,1). For plot legend, refer next figure.
APPENDIX B. SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Figure B.4: Values for Nash Sutcliffe Efficiency (NSE) and Kling Gupta Efficiency (KGE) for different combinations of standard deviation $\sigma$ and bias $\beta$. Linear correlation $r$ is constant at 0.9. The analysis is done for two values of standard deviation: $\sigma = 0.7$ mm/d (10th percentile - 9 catchments of 86 had a lower value); and $\sigma = 4.0$ mm/d (90th percentile - 9 catchments of 86 had a higher value). Note that Equation 4 from Gupta et al. (2009) requires the average flow to be defined in absolute terms; a value of 0.7 mm/d or 255 mm/yr was assumed, which was close to the median value for the study catchments. Orange and purple values mark issues with the NSE, as follows: Orange: Gupta et al. (2009) noted that the bias term is normalised by the observed standard deviation, which means that in catchments with high flow variability (as in this study) the magnitude of the bias can be high without penalising the NSE score. Purple: Gupta et al. (2009) noted a further problem with the treatment of the standard deviation $\sigma$ in the NSE, regarding the ratio $\sigma_{\text{simulated}} / \sigma_{\text{observed}}$. Although this ratio should ideally have a value of unity, the optimum value for NSE occurs when the ratio is equal to the linear correlation.
Figure B.5: Pareto curves, for each of the five model structures, in 10 selected study catchments, along with single-objective optimization results using CMA-ES.
Figure B.6: Four cases where the GR4J and GR4JMOD curves have similar endpoints. In each case the single-objective Differential Split Sample Test results would lead to the erroneous conclusion that the improvements introduced into GR4JMOD by Hughes et al. (2013) made negligible difference to the model’s capabilities. (a) 401210 Snowy Creek below Granite Flat, Victoria; (b) 405217 Yea River at Devlins Bridge, Victoria; (c) 143110A Bremer River at Adams Bridge, Queensland; (d) A5040517 First Creek at Waterfall Gully, South Australia.
Figure B.7: Map of catchments where no model structure could meet a given modelling standard.
Figure B.8: Long-term simulation bias for selected parameter sets, after Coron et al. (2014, Figure 5), 1 of 4. Simulation bias is plotted as a ten-year moving average for selected parameter sets from the Pareto curve. The ten-year average streamflows are also plotted for reference, in blue. This figure is the first of a series of four, showing a total of 21 case studies selected based on the shape of their Pareto Fronts. Specifically, three criteria were used to select these 21 case studies: (1) At least 1 parameter set within \([0.8^+, 0.8^+]\) - that is, the model structure meets Standard 2 in this catchment (refer Chapter 4 for definitions); (2) Calibrating to the non-dry period in isolation results in reductions in \(\text{KGE}_{\text{dry}}\) by at least 0.10 (ie. to 0.70) relative to the parameter set in (1); and (3) Calibrating to the dry period in isolation results in reductions in \(\text{KGE}_{\text{non-dry}}\) by at least 0.10 (ie. to 0.70) relative to the parameter set in (1).
Figure B.9: Long-term simulation bias for selected parameter sets, after Coron et al. (2014, Figure 5), 2 of 4.
Figure B.10: Long-term simulation bias for selected parameter sets, after Coron et al. (2014, Figure 5)), 3 of 4.
Figure B.11: Long-term simulation bias for selected parameter sets, after Coron et al. (2014, Figure 5), 4 of 4.
Appendix C

Supplementary material for Chapter 5

C.1 Introduction

This appendix contains the following:

1. Combined ECDFs for each model structure (Section C.2) for hydrologic signatures (Figure C.1) and objective functions (Figure C.2) with accompanying explanation.

2. Method description (Section C.3) describing steps to accommodate the two-metric ‘Flow Duration Curve’ signature.

3. Selected figures and tables (Section C.4):

   (a) List of case studies (Table C.1) accepted for analysis as per the rules outlined in Sections 5.3.4 and 5.3.5.

   (b) Histograms of typical values for hydrologic signatures observed over a large set of Australian catchments (Figure C.3).

   (c) Example plots showing Pareto Fronts and LHS random samples in 2D objective space (Figure C.4).
C.2 Combined ECDFs for each model structure

Figures C.1 and C.2 are based on the same data as Figures 5.3 and 5.4, but formulated in a different way. Figure C.1 is intended to show that the ECDFs of certain signatures had the properties described in the Thesis (Section 5.4.1):

There were some cases where the adjusted KSD value seemed to show promise, but the actual ECDFs indicated otherwise. This was particularly the case if both regions enveloped the observed value but showed a very wide range of signature values. This problem affected ... [signature selection, with some signatures being] ... removed from the list after manual inspection of the ECDFs revealed this problem. The removed signatures were High Spell Duration $C_0$ (SIMHYD, 1HACRES) and High Spell Count (1HACRES, SACRAMENTO).

Ideally, the best plot to show to demonstrate this problem would be the ECDF for each case study affected by the problem. However, this would mean showing dozens of ECDF plots here. Instead, Figures S1 and S2 show ECDFs based on a combination of every ECDF plot for a given signature and model structure. There is more than one way to combine ECDF ordinates; in this case the following procedure was used:

1. For hydrologic signatures only, values were standardized in the following way. Firstly, the signature error was calculated and used in place of the actual signature value (in contrast to Figure 5.1). The typical variability in observed signature value across catchments was quantified by taking the interquartile range of values observed in a large set of catchments (see below). The signature error was then standardized by dividing by this interquartile range.

2. Secondly, ecdfs were combined so that for each y value, the average of all x values was plotted. This tended to average out inter-catchment variability. Even after this removal of variability, it is still seen that, in the cases listed above, the ECDFs exhibit the stated behavior, namely 'both regions enveloped the observed value but showed a very wide range of signature values'.
Figure C.1: Combined ECDFs for each model structure, for hydrologic signatures. See Section C.2 text for explanation. Signatures selected for each model structure are circled in red.

The 'large set of catchments' mentioned in (1) was the 221 Hydrologic Reference Stations (HRS) of Australia’s Bureau of Meteorology. These catchments are from around Australia, including tropical, temperate and arid areas (Turner, 2012), from which the 86 study catchments are selected. For completeness, below we provide histograms of the 221 observed values for each hydrologic signature (Figure C.3).
Figure C.2: Combined ECDFs for each model structure, for objective functions. See Section C.2 text for explanation.

C.3 Method description for two-metric ‘Flow Duration Curve’ signature

As discussed in Section 5.3.2, the signature “flow duration curve” is different from every other hydrologic signature because it is a two-metric signature. It is composed of two individual metrics, FDC1 and FDC2, as explained by Vrugt and Sadegh (2013). In the context of signature replication, if model simulations produce a match with FDC1, this is meaningless unless FDC2 is matched also (and vice versa). This means that the data in Figures 1 and 4 would likewise be meaningless (for FDC1 and 2) unless action was taken to account for the dual-metric nature of this signature.

Furthermore, FDC1 and 2 are not actually observed quantities; they are themselves parameters in a parametric representation of the flow duration curve (cf. Sadegh and Vrugt, 2014). In other words, FDC1 and 2 are ‘optimum’ parameters such that the parametric equation provides the best possible fit to the observed flow duration curve, as explained in Vrugt and Sadegh (2013). This is why a match with FDC1 is meaningless unless FDC2 is matched too (and vice versa).

For Chapter 5 only, the procedure for accounting for the dual nature of the flow duration curve signature was:
C.4 SELECTED FIGURES AND TABLES

a. Fit the parametric form to the observed flow duration curve, determining FDC1_{obs} and FDC2_{obs};

b. For FDC1, hold FDC2 constant at FDC2_{obs}, and then fit the parametric form to the simulated flow by varying FDC1 only. The value that provides the best fit is the ‘simulated’ value of FDC1.

c. For FDC2, follow the same procedure as (b), with FDC1 and FDC2 reversed.

To complete the explanation, we provide the following information on the fitting process. To fit the equation to the observed flow duration, we considered the match at 10 discrete points, namely the 5th, 15th […] 95th percentiles. An optimiser was used to find the parameter value(s) that provided the smallest sum of absolute errors across these ten points. While the parametric form tended to provide a good fit for perennial catchments, ephemeral catchments were not well fitted in the low flow (<0.01mm/d) and zero flow parts of the flow duration curve. This is not surprising since the parametric form is only 2 parameters, so it has limited flexibility. To deal with this problem, only ordinates with flow values above 0.01 mm/d were included in the sum of absolute errors. Since the lowest flow values were thus ignored, the ‘flow duration curve’ signature should not be considered to provide information about the low-flow or cease-to-flow behavior of a catchment.

C.4 Selected figures and tables
Table C.1: List of case studies accepted for analysis as per the rules outlined in Sections 5.3.4 and 5.3.5. The 6 case studies rejected under Section 5.3.4 were: IHACRES: 206014, 206018, and 407214; GR4JM OD: 416008; 418014; 136203. State and territory abbreviations are: ACT - Australian Capital Territory; NSW - New South Wales; QLD - Queensland; SA - South Australia; VIC - Victoria; WA - Western Australia. This list shows that all states and territories within the study area were represented in the selected set for each model, with the exception of the ACT, which is relatively small with few catchments.

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>GR4J</th>
<th>SIMH</th>
<th>IHAC</th>
<th>GR4JM</th>
<th>SACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cotter River at Gingera (410730)</td>
<td>ACT</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Wollomombi River at Coninside (206014)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Apsley River at Apsley Falls (206018)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Nowendoc River at Nowendoc (206007)</td>
<td>NSW</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Barnard River at Barry (208009)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>William River at Tillegra (210011)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Jigadee Creek at Avondale (211008)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Kowmung River at Cedar Ford (212260)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Clyde River at Brooman (216002)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Currambene Creek at Falls Creek (216004)</td>
<td>NSW</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Bowna Creek at Yambra (401015)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Goobarragandra River at Lacmalac (410057)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Adelong Creek at Batlow Road (410061)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Tenterfield Creek at Clifton (416003)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Beadt River at Haystack (416008)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Copes Creek at Kimberley (418005)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Gwydir River at Yarrowyck (418014)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Barambah Creek at Litzows (136202D)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Barkler Creek at Brooklands (136203A)</td>
<td>QLD</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Gregory River at Burrum Highway (137101A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Isis River at Bruce Highway (137201A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>
C.4. SELECTED FIGURES AND TABLES

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>GR4J</th>
<th>SIMH</th>
<th>IHAC</th>
<th>GR4JM</th>
<th>SACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tinana Creek at Tagigan Road (138009A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Wide Bay Creek at Killayan (138010A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Kandanga Creek at Hygait (138113A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Bremer River at Adams Bridge (143110A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Stanley River at Peacher (143303A)</td>
<td>QLD</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Teviot Brook at Croftby (145011A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Burnett Creek at Up Stream Maroon Dam (145018A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Coomera River at Army Camp (146010A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Currumbin Creek at Nicolls Bridge (146012A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Back Creek at Beechmont (146014A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tallebudgera Creek at Tallebudgera Ck Rd (146055A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Swan Creek at Swanfels (422306A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Spring Creek at Killarney (422321B)</td>
<td>QLD</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Kings Creek at Aides Bridge (122334A)</td>
<td>QLD</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Scott Creek at Scott Bottom (A5030502)</td>
<td>SA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>First Creek at Waterfall Gully (A5040517)</td>
<td>SA</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Sixth Creek at Castambul (A5040523)</td>
<td>SA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>North Para River at Penrice (A5050517)</td>
<td>SA</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Rocky River upstream Gorge Falls (A5130501)</td>
<td>SA</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Errimundra River at Errinundra (221207)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Wonnangatta River at Crooked River (224206)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Macalister River at Glencairn (225219)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Latrobe River at Near Noojee (226222)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tarwin River East Branch at Dumbalk North (227226)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Wilkura Creek at Leongatha (227227)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Arkins Creek West Branch at Weyangta (235205)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Cudgewa Creek at Berringama (401208)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Name</td>
<td>State</td>
<td>GR4J</td>
<td>SIMH</td>
<td>IHAC</td>
<td>GR4JM</td>
<td>SACR</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Snowy Creek at Below Granite Flat (401210)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Nariel Creek at Upper Nariel (401212)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Gibbo River at Gibbo Park (401217)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Yackandandah Creek at Osbornes Flat (402204)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Running Creek at Running Creek (402206)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Flaggy Creek at Myrtleford Road Bridge (402217)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Happy Valley Creek at Rosewhite (403214)</td>
<td>VIC</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Rose River at Matong North (403217)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Buffalo River at Abbeyard (403222)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Boggy Creek at Angleside (403226)</td>
<td>VIC</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Murrindindi River at Murrindindi above Colwells (405205)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Acheron River at Taggerty (405209)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Howqua River at Glen Esk (405215)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Yea River at Devlins Bridge (405217)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Jamieson River at Gerrang Bridge (405218)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Goulburn River at Doherty (405219)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Big River at Jamieson (405227)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Mollison Creek at Pyalong (405238)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Major Creek at Graytown (405248)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Home Creek at Yareck (405274)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Campaspe River at Ashborne (406208)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Campaspe River at Redesdale (406213)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Axe Creek at Longlea (406214)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Mount Pleasant Creek at Runnymede (406224)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Creswick Creek at Clunes (407214)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Bet Bet Creek at Norwood (407220)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Joyces Creek at Strathlea (407230)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Piccaninny Creek at Minto (407253)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Name</td>
<td>State</td>
<td>GR4J</td>
<td>SIMH</td>
<td>IHAC</td>
<td>GR4JM</td>
<td>SACR</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Avoca River at Amphitheatre (408202)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Richardson River at Carrs Plains (415226)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Concongella Creek at Stawell (415237)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tarra River at Fishers (227225A)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Fifteen Mile Creek at Greta South (403213A)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Shannon River at Dog Pool (606185)</td>
<td>WA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Dombakup Brook at Malimup Track (607155)</td>
<td>WA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Carey Brook at Staircase Road (608002)</td>
<td>WA</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Harvey River at Dingo Road (613002)</td>
<td>WA</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Clarke Brook at Hillview Farm (613146)</td>
<td>WA</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>18</td>
<td>18</td>
<td>21</td>
<td>28</td>
<td>26</td>
</tr>
</tbody>
</table>
APPENDIX C. SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Figure C.3: Histograms of values of hydrologic signatures in observed flow for each of the 221 Hydrologic Reference Stations of Australia’s Bureau of Meteorology Turner and Van Zandt (2012). Note that the large number of values at the upper end of the range for ‘Skew daily’ and ‘IQR ratio’ (both of which are fractions in form) are the result of the author’s choice to assign a value of 1.0 and 100.0, respectively, whenever the denominator was zero. For units, refer to Table 5.2.
Figure C.4: Selection of Pareto fronts (in red) and random samples (in blue) generated for Chapter 5. Case studies on the left experienced no problems populating the region near the Pareto Front; whereas case studies on the right showed poor sampling density in the most robust regions of the Parameter Space.
Appendix D

Supplementary material for
Chapter 6

D.1 Introduction

This appendix contains the following:

1. **DREAM-ABC settings (with associated analysis).** Relating to Section 6.3.3, this section provides the rationale for the adopted values of the following DREAM-ABC settings: ε value; number of chains; number of behavioral chains required to characterize the behavioural space; and minimum behavioural population of each behavioural chain.

2. **Prior study: search for signature combinations leading to high KGE.** Relating to Section 6.3.3, this summary includes rationale, methods, results and discussion of a previously unpublished study. The results of this prior study were used to inform which signatures were combined with shortlisted signatures to screen out parameter sets that performed poorly over the nondry (calibration) period, as described in Section 6.3.3.

3. **Selection rules for choosing subsets of catchments for DREAM-ABC testing**, along with the resulting lists. These rules are referenced in Section 6.3.3.
4. **Selected hybrid method results expanding on Section 6.4.3**, showing the same information as Figure 6.2 for a wider range of catchments.

5. **Data depth analysis**, showing the performance of parameter sets with high depth compared to the median values of their parent ensembles, as referenced in Sections 6.3.3 and 6.3.3.

6. **Repeat of Figure 6.1 results, using alternative evaluation metrics.** Referenced in Sections 6.4 and 6.5.

### D.2 DREAM-ABC settings (with associated analysis)

#### D.2.1 Overview of DREAM-ABC settings

Using the DREAM ABC algorithm required numerous decisions regarding the technical aspects of the search method. Generally, we followed examples given by the authors (*Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014; Vrugt, 2016*) with some modifications to ease the computational burden presented by the 3000 runs of DREAM-ABC undertaken for Chapter 6. The following paragraphs summarise the main DREAM-ABC settings.

**Number of chains:** 10 Markov Chains were used for all model structures except for SACRAMENTO. This was consistent with the ABC example given by (*Vrugt, 2016, Section 5.4*) who used 10 chains for the 7-parameter HYMOD. For SACRAMENTO, d+1 chains (14) were used, where d means dimensionality - that is, number of free model parameters. This was consistent with the example for this model structure given by (*Sadegh and Vrugt, 2014, p.6780*).

**Acceptance of parameter sets:** following the comments of (*Vrugt and Sadegh, 2013, p6774*), we discarded as non-behavioral any parameter set that was not within a certain distance, ε, of any signature in the combination in question. Adopted ε values are discussed in the following paragraph. All results in Chapter 6 display the behavioral solutions only - note that this is different from *Vrugt and Sadegh* (2013) who displayed non-behavioral solutions as part of their discussion of
a case study where the model was “unable to simultaneously satisfy all the four different summary metrics used herein” (p6784). In the Chapter 6, such a case study would have been characterised as a model failure and labelled as such in the results. We capped the number of function evaluations to 200,000, the same number used by Sadegh and Vrugt (2014), except in certain circumstances where runs were allowed to continue up to 500,000 (Section D.2.2).

**Degree of closeness (ε) in signature matching:** given differences in the magnitudes of each signature - for example, slow flow proportions vary between 0 and 1 whereas median annual flows vary between 0 and 3000mm/year - it was logical to adopt a value of ε that was relative rather than absolute. Based on the results described below (Section D.2.2) we adopted ε = 2%. That is, a parameter set was accepted as behavioral if signature values were within 2% of the values in the observed data. It is noted that observed signatures values are themselves uncertain (eg, Westerberg et al., 2016), but we did not alter the DREAM-ABC framework to account for uncertainty in hydrologic signatures, and note this as a potential future research area.

**Number and population of behavioural chains:** it is possible that some chains may find behavioral parameter sets while others may not. Although in theory even a single behavioral chain may provide good coverage of the behavioral parameter space, in practice it was found that requiring multiple behavioral chains (5), and requiring that each chain had a minimum number of behavioral parameter sets (150) provided a more robust basis for generating an ensemble, as discussed below.

### D.2.2 Analysis supporting adopted settings

In this section we summarise analysis supporting each of the points noted above. This section presents the points in reverse order to the previous section.

**Population of behavioral chains:** in standard DREAM, which does not have the binary behavioral/non-behavioral nature of ABC, all parameter sets on all chains are considered equally valid, except for those at the start of the run (the ‘burn-in’ period) which are discarded (Vrugt et al., 2008). The run is continued until
the chains have converged on the posterior distribution and are thus mixing with one another in the same parameter space, as indicated by a Gelman-Rubin statistic value less than 1.2 (Gelman and Rubin, 1992). In DREAM-ABC ideally the run should continue at least until the behavioral parts of the chains attain a Gelman Rubin value less than 1.2, indicating that the chains have thoroughly searched the behavioral space. However, we found that a Gelman Rubin statistic less than 1.2 was not possible in some cases, even after five million function evaluations (Figure D.1), meaning that this was not a sufficient basis for run termination. Across the 17 case studies tested (case studies are listed in Section D.3), the Gelman-Rubin statistic tended to converge (i.e. the behavior shown in Figure D.1) for behavioral chain populations of 150 or more. Together with the information in the following section, this assisted in formulating the rules listed below.

Number of behavioral chains: because the chains were not always fully mixing in the behavioral space, having more behavioral chains will likely ensure a more thorough search of the behavioral space. Based on a bootstrap analysis (Figure D.2) we found that, across the 17 case studies listed in Appendix D.3, a requirement of 5 behavioral chains was generally sufficient to ensure that the parameter variability across all chains was 90% of that attained with 20 behavioral chains.

Rules: based on the above, the following rules were adopted for DREAM-ABC:

- 10 Markov Chains were used for all model structures except for SACRAMENTO which used 14, as mentioned above;

- At least 5 behavioral chains, each with 150 behavioral parameter sets, were required for the run to be considered behavioral;

- If a run of DREAM-ABC reached 200,000 function evaluations (Sadegh and Vrugt, 2014) without finding any behavioral parameter sets, it was a model failure.

- If a run of DREAM-ABC successfully identified behavioral parameter sets, it was terminated at 200,000 function evaluations unless the above rules were not
yet met (5 chains, 150 behavioral sets) in which case the run continued until the rules were met or 500,000 function evaluations were complete.

**Degree of closeness (ε) in signature matching:** for the reasons outlined in the previous section, the ε value was defined in relative rather than absolute terms. As per the method demonstrated in Figure 4 of Vrugt and Sadiq (2013), one way of choosing the ε value is to trial multiple ε values and to observe at what value the parameter variance of the behavioral ensemble stabilises. Beyond this point, the expected behavior was that “lower values of ε provide similar posterior estimates, yet unnecessarily increase the computational burden of the ABC analysis.” (ibid p4339). This was conducted for the 17 case studies listed in the following section, for two signature combinations: mean & flow duration curve & slow flow (0.925) (cf. Vrugt and Sadiq, 2013); and mean & C2 daily & catchment lag A. Example results are shown in Figure D.3. The boxplots show that the behavioral ensemble shrunk in the parameter space as ε moved through values 0.2, 0.1, 0.05 and 0.02; the higher accuracy requirements also led to a lower success rate, and thus lower behavioral populations in each chain (indicated by colour). For values less than 0.02, the behavioral population continued to shrink, and each individual chain only explored a portion of the behavioral space, so that the parameter values were not consistent between chains. Using a lower ε value than 0.02 did not result in a smaller behavioral parameter space. Based on these results (and supporting results from other case studies - not shown), an ε value of 0.02 was adopted in Chapter 6.

**D.3 Prior study: signature combinations leading to high KGE**

Prior to commencement of research for Chapter 5, a (previously unpublished) study was undertaken, as summarised below. This summary is provided because the findings of the prior study influenced which signature combinations were used in DREAM-ABC analysis to screen out parameter sets with a low KGE over the nondry (calibration) period, as described in Section 6.3.3 of Chapter 6.
Figure D.1: Gelman Rubin statistic as a function of behavioral population in each DREAM-ABC chain, for three different case studies. The statistic is calculated based on the behavioral portions of chains only. Each case study is based on a model structure applied in a single catchment. Two of the three case studies do not achieve a Gelman-Rubin statistic value of 1.2 (marked as red dashes). Runs of DREAM-ABC generating 500 behavioral sets in each chain often required >1,000,000 function evaluations.

**Hypothesis and aim:** The hypothesis of the study was that there exists a set of signatures that, when matched in an ABC framework, identify parameter sets that provide robust simulation with changing climate, regardless of which model structure is used. The key difference with the DREAM-ABC analysis in Chapter 6 was the prior study searched for a model-independent set of signatures, i.e. one that would work well for any model structure. The aim of the study was to find this signature set, should it exist. In contrast, the Chapter 6 accepted a priori that signatures must be chosen on a model-by-model basis, an assumption based on the joint results of the prior study and the Exploratory Analysis described in 5.

**Method:** The method was based on the DREAM-ABC algorithm, using the same settings as used in Chapter 6, as described in Appendix D.2. The signatures
Figure D.2: Effect of number of chains on the volume of parameter space occupied by behavioral parameter sets (as represented by the C_v of parameter values), for the same case studies as above. Each case study is based on a model structure applied in a single catchment. For two of the three case studies, two behavioral chains were sufficient to sample the same volume as 20; for the third (Case 3) a larger number was required.

were chosen in a stepwise approach analogous to stepwise linear regression, as described in the following points, which were undertaken for each of 17 case studies. Case study selection is described at the end of the method description.

1. For each of the 35 hydrologic signatures (Chapter 5, Table 5.2):

   (a) Conduct runs of DREAM-ABC matching to the individual signature, and no others.

   (b) Use the results of 1a to determine the signature that provides the highest KGE value**, on average over all 17 case studies. Call this ‘signature 1’.

2. For each of the 34 remaining signatures:

   (a) Conduct a run of DREAM-ABC matching to this signature and signature
Figure D.3: Effect of $\epsilon$ value on the volume of parameter space occupied by behavioral parameter sets in 10 chains, for a single GR4J case study, based on five million function evaluations. Lower $\epsilon$ value reduced the success rate dramatically, as evidenced by the chain populations, and higher variability due to smaller sample size caused the consistency between chains to decrease for $\epsilon < 0.02$.

1. (b) Use the results of 2a to determine the signature that provides the highest KGE value**, when used together with signature 1, on average over all 17 case studies. Call this ‘signature 2’. In this step, signatures with separate information content to signature 1 will tend to be chosen over those that share information content with signature 1.

3. For each of the 33 remaining signatures...

   (and so on, until the model structures fail or no further improvement is possible)

** over either the nondry (calibration) period or the dry (evaluation) period, depending on the stage of research, as explained below.

As with Chapter 6, all signature matching was undertaken over the nondry (calibration) period, not the dry (evaluation) period.

The set of case studies was set at 17 to make the analysis feasible, since the above steps require so many DREAM runs ($36+35+34+\ldots$). The criteria used were:
D.3. PRIOR STUDY OF SIGNATURE COMBINATIONS

- Rule 1: same as Rule 1 in Section 5.3.4, except that a KGE threshold of 0.8 was used instead of 0.7;

- Rule 2: same as Rule 2 in Section 5.3.4, except that a fixed threshold, not a relative threshold, was used: the $KGE_{dry}$ obtained by the reference method had to be less than 0.6.

The 17 case studies covered all five model structures and were relatively well spread geographically between the different states of Australia.

**Results and discussion:** The first stage of research was focused on maximising KGE in the nondry (calibration) period. The results (Table D.1) indicated that that pair of signatures providing highest $KGE_{nondry}$ were the mean and the $cv$ daily. This was not surprising given the central place these signatures have in KGE formulation (Gupta et al., 2009).

When choosing signature 3, Catchment Lag A and Slow Flow 0.925 were difficult to choose between. Both of the following signature combinations led to a high KGE in the nondry (calibration) period:

1. Mean daily flow & $C_v$ daily & Catchment Lag A; and

2. Mean daily flow & $C_v$ daily & Slow flow 0.925

This is why they were both chosen for inclusion in Section 6.3.3. When Catchment Lag A was selected as signature 3, Slow Flow 0.925 was subsequently selected as signature 4, indicating that the Catchment Lag A and the Slow Flow 0.925 are likely to contain complementary (rather than shared) information content. This is why the combination Mean daily & $C_v$ daily & Catchment Lag A & Slow Flow 0.925 was also used in Section 6.3.3.

Thus ends the part of this summary relevant to the Chapter 6. For completeness, we also provide the results of the next stage. Rather than focusing on $KGE_{nondry}$, the focus of the next stage was $KGE_{dry}$. The signatures were still matched over the nondry (calibration) period, but the basis of signature selection was maximization of $KGE_{dry}$. 
Table D.1: Prior study results, first stage. Each successive signature was chosen to maximize the KGE in the calibration (nondry) period. Median KGEs quoted were calculated based on 17 case studies. Signature combinations in this table influenced signature selection in Chapter 6 (Section 6.3.3)

<table>
<thead>
<tr>
<th>Signature 1</th>
<th>Signature 2</th>
<th>Signature 3</th>
<th>Signature 4</th>
<th>Signature 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean daily flow (MDF)</td>
<td>$C_v$ daily (CVD)</td>
<td>Catchment Lag A (CLA)</td>
<td>Slow Flow 0.925 (SFA)</td>
<td>None selected (no benefit)</td>
</tr>
<tr>
<td>Combination</td>
<td>MDF only</td>
<td>MDF &amp; CVD</td>
<td>MDF &amp; CVD &amp; CLA</td>
<td>MDF &amp; CVD &amp; CLA &amp; SFA</td>
</tr>
<tr>
<td>Performance for this combination*</td>
<td>0.25</td>
<td>0.76</td>
<td>0.82</td>
<td>0.84</td>
</tr>
</tbody>
</table>

* Median KGE in calibration (nondry) period

This change in focus yielded a different signature set (Table D.2) to the first stage. The first signature was the Flow Duration Curve (FDC). Matching to FDC provided a higher $KGE_{dry}$ than any two-signature combination tested. Subsequent steps added the Slow Flow 0.925 signature (the only one in common with the previous stage) and the Wetting Up Proportion of Years.

However, as seen in Table D.2, the $KGE_{dry}$ values obtained were not high, particularly considering that each case study was capable of $KGE_{dry} > 0.8$ as per the selection rules. The $KGE_{dry}$ values for the combinations listed in Table D.1 were also relatively low. Thus, the overall conclusion was that the hypothesis is false; i.e. that there is no single, model-independent combination of signatures that provides robust model performance in changing climate. This conclusion encouraged consideration of model-specific signatures which led to the Exploratory Analysis described in Chapter 5.

### D.4 DREAM-ABC case study selection rules

As explained in Section 6.3.3, DREAM-ABC analysis involves testing the calibration potential of combinations of hydrologic signatures. With four shortlisted signatures per model structure from Chapter 5, each combined with three possibilities from
Table D.2: Prior study results, second stage. Each successive signature was chosen to maximize the KGE in the evaluation (dry) period. Median KGEs quoted were calculated based on 17 case studies.

<table>
<thead>
<tr>
<th>Selected Signature</th>
<th>Signature 1</th>
<th>Signature 2</th>
<th>Signature 3</th>
<th>Signature 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>Flow duration curve (FDC)</td>
<td>Slow flow 0.925 (SFA)</td>
<td>Wetting up proportion of years (WUPY)</td>
<td>None selected (no benefit)</td>
</tr>
<tr>
<td>Performance for this combination*</td>
<td>0.32</td>
<td>0.56</td>
<td>0.62</td>
<td>-</td>
</tr>
</tbody>
</table>

* Median KGE in evaluation (dry) period

The previous section, there were many combinations to be tested. Since the computational requirements of each run of DREAM-ABC are relatively high (Appendix D.2), it was decided to reduce the set of catchments on which DREAM-ABC would be tested using the following rules to 12 catchments per model structure. Note: the combinations with highest performance for each model structure among the subsample of 12 catchments were then run for all 86 catchments, as mentioned in Section 6.4.4 (cf. Figure 6.6).

The rationale and rules of selection were identical to those used in 6.4.4 of Chapter 5, with the addition of a random selection step at the end. As with Chapter 5, the logic was to focus on catchments where models performed well during wetter periods but poorly during drier periods, since this is the behaviour that prompted this study. Furthermore, it was useful to ignore case studies where the Pareto analysis of Chapter 4 indicated that no parameter set exists that attains a high KGE value over both nondry and dry periods. Since no parameter sets exist with the desired behaviour, improving such case studies is beyond the scope of any calibration method, no matter how good. The same two selection rules as Chapter 5 were adopted, namely:

1. Case studies should be capable of KGE>0.7 in both the nondry (calibration) period and the dry (evaluation) period, according to the Pareto analysis undertaken by Chapter 4.
2. Case studies were selected where optimisation to KGE over the nondry (calibration) period resulted in a KGE in the dry (evaluation) period that was lower by at least 0.2 than the parameter set marked in Figure 5.1 of Chapter 5 as “closest to [1, 1]”.

Taken together, these rules tended to select only those case studies where high KGE performance was possible over the two different climatic conditions tested, but the reference method failed to select a parameter set that provided this robust performance.

As described in Chapter 5, of the set of 86 catchments, this process selected subsets of between 18 (in the case of GR4J) and 33 (in the case of SACRAMENTO) catchments. Then, to reduce the number further, random selection was used to select a set of 12 catchments for each model structure.

The end result of these rules is shown in Table D.3 below.

Table D.3: List of case studies accepted for analysis as per the rules outlined above. Cases with an asterisk mean that rules 1 and 2 described in Section D.3 were fulfilled but the case study was not randomly selected. State abbreviations are: ACT - Australian Capital Territory; NSW - New South Wales; QLD - Queensland; SA - South Australia; VIC - Victoria; WA - Western Australia.

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>GR4J</th>
<th>SIMH</th>
<th>IHAC</th>
<th>GR4JM</th>
<th>SACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cotter River at Gingera (410730)</td>
<td>ACT</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Wollomombi River at Coninside (206014)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Apsley River at Apsley Falls (206018)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Nowendoc River at Nowendoc (208007)</td>
<td>NSW</td>
<td>y</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Barnard River at Barry (208009)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>William River at Tillega (210011)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Jigadee Creek at Avondale (211006)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Kowmung River at Cedar Ford (212060)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Clyde River at Brooman (216002)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Currumbene Creek at Falls Creek (216004)</td>
<td>NSW</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Bowna Creek at Yamba (401015)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Name</td>
<td>State</td>
<td>GR4</td>
<td>SIMH</td>
<td>IHAC</td>
<td>GR4JM</td>
<td>SACR</td>
</tr>
<tr>
<td>-----------------------------------------------------------</td>
<td>-------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Goobarragandra River at Lacmalac (410057)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Adelong Creek at Batlow Road (410061)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Tenterfield Creek at Clifton (416003)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n*</td>
</tr>
<tr>
<td>Bead River at Haystack (416008)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Copes Creek at Kimbeley (418005)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Gwydir River at Yarrowyck (418014)</td>
<td>NSW</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Barambah Creek at Litzows (136302D)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Barker Creek at Brooklands (136303A)</td>
<td>QLD</td>
<td>n</td>
<td>y</td>
<td>n*</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Gregory River at Burnum Highway (137101A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Isis River at Bruce Highway (137201A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Timna Creek at Tagigan Road (138009A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
</tr>
<tr>
<td>Wide Bay Creek at Killian (138010A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Kandanga Creek at Hygait (138113A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Bremer River at Adams Bridge (143110A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
<td>n*</td>
</tr>
<tr>
<td>Stanley River at Pecheister (143303A)</td>
<td>QLD</td>
<td>n*</td>
<td>n*</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Teviot Brook at Croftby (145011A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Burnett Creek at Up Stream Maroon Dam (145018A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Coomera River at Army Camp (146010A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Currumbin Creek at Nicolls Bridge (146012A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Back Creek at Beechmont (146014A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tallebudgera Creek at Tallebudgera Ck Rd (146005A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Swan Creek at Swanfels (422306A)</td>
<td>QLD</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Spring Creek at Killarney (122321B)</td>
<td>QLD</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
</tr>
<tr>
<td>Kings Creek at Aides Bridge (122334A)</td>
<td>QLD</td>
<td>n*</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>y</td>
</tr>
<tr>
<td>Scott Creek at Scott Bottom (A503002)</td>
<td>SA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>First Creek at Waterfall Gully (A5040517)</td>
<td>SA</td>
<td>y</td>
<td>n*</td>
<td>n*</td>
<td>y</td>
<td>n*</td>
</tr>
<tr>
<td>Sixth Creek at Castambul (A5040523)</td>
<td>SA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>North Para River at Penrice (A5050517)</td>
<td>SA</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
</tbody>
</table>
## APPENDIX D. SUPPLEMENTARY MATERIAL FOR CHAPTER 6

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>GR4J</th>
<th>SIMH</th>
<th>IHAC</th>
<th>GR4JM</th>
<th>SACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rocky River upstream Gorge Falls</td>
<td>SA</td>
<td>n*</td>
<td>n</td>
<td>y</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>(A5130501)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Errinundra River at Errinundra (221207)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Wonnangatta River at Crooked River (224206)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Macalister River at Glencairn (225219)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Latrobe River at Near Noojee (236222)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tarwin River East Branch at Dumbalk North (227226)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Willur Creek at Leongatha (227227)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Arkins Creek West Branch at Wyelangta (235205)</td>
<td>VIC</td>
<td>n</td>
<td>n*</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Cudgewa Creek at Berringama (401208)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n*</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Snowy Creek at Below Granite Flat (401210)</td>
<td>VIC</td>
<td>n*</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Nariel Creek at Upper Nariel (401212)</td>
<td>VIC</td>
<td>y</td>
<td>n*</td>
<td>n*</td>
<td>n*</td>
<td>n*</td>
</tr>
<tr>
<td>Gibbo River at Gibbo Park (401217)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Yackandandah Creek at Osbornes Flat (402204)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Running Creek at Running Creek (402206)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Flaggy Creek at Myrtleford Road Bridge (402217)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>y</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Happy Valley Creek at Rosewhite (403214)</td>
<td>VIC</td>
<td>y</td>
<td>n</td>
<td>n*</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Rose River at Matong North (403217)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Buffalo River at Abbeyard (403222)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Boggy Creek at Angleside (403236)</td>
<td>VIC</td>
<td>n*</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>n</td>
</tr>
<tr>
<td>Murrindindi River at Murrindindi above Coles (405205)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Acheron River at Taggerty (405209)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Howqua River at Glen Esk (405215)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Yea River at Devlins Bridge (405217)</td>
<td>VIC</td>
<td>y</td>
<td>n*</td>
<td>y</td>
<td>y</td>
<td>n*</td>
</tr>
<tr>
<td>Jamieson River at Gerrang Bridge (405218)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Goulburn River at Dohertys (405219)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
</tbody>
</table>
## D.5 Hybrid method expanded results

In this section, we provide Hybrid method results like those shown in Figure 6.2, but for all 86 catchments, for two selected cases. In Figure D.4, results across all catchments are shown for the signature $C_v$ annual and the model structure GR4J. In Figure D.5, results are shown for the signature High Flow Discharge and the model

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>GR4J</th>
<th>SIMH</th>
<th>IHAC</th>
<th>GR4JM</th>
<th>SACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big River at Jamieson (405227)</td>
<td>VIC</td>
<td>n</td>
<td>n*</td>
<td>n*</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Mollison Creek at Pyalong (405238)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>Major Creek at Graytown (405248)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Home Creek at Yarck (405274)</td>
<td>VIC</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n*</td>
<td>y</td>
</tr>
<tr>
<td>Campaspe River at Ashborne (406208)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Campaspe River at Redesdale (406213)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Axe Creek at Longlea (406214)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Mount Pleasant Creek at Runnymede (406224)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Creswick Creek at Chunes (407214)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Bet Bet Creek at Norwood (407220)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Joycees Creek at Strathlene (407230)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n*</td>
</tr>
<tr>
<td>Piccaninny Creek at Minto (407253)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Avoca River at Amphitheatre (408202)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Richardson River at Carrs Plains (415226)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Concongella Creek at Stawell (415237)</td>
<td>VIC</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Tarra River at Fischers (227225A)</td>
<td>VIC</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Fifteen Mile Creek at Greta South (403213A)</td>
<td>VIC</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Shannon River at Dog Pool (606185)</td>
<td>WA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Dombaleup Brook at Malimup Track (607155)</td>
<td>WA</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Carey Brook at Staircase Road (608002)</td>
<td>WA</td>
<td>n*</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Harvey River at Dingo Road (613002)</td>
<td>WA</td>
<td>y</td>
<td>n*</td>
<td>n*</td>
<td>n*</td>
<td>n*</td>
</tr>
<tr>
<td>Clarke Brook at Hillview Farm (613146)</td>
<td>WA</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
</tbody>
</table>

| TOTAL                                                                 | 12    | 12   | 12   | 12   | 12    |

---
structure SIMHYD.

Chapter 5 results indicated strong potential for both of these signatures. However, the full results shown here indicate that the gains in model performance from the hybrid method, while generally favourable across the subsample of catchments used in Chapter 5 (marked with grey boxes), results across the full set of catchments were relatively less favourable.

D.6 Do deeper parameter sets perform better in split sample testing?

Bárdossy and Singh (2008) found that, in ensembles of well-performing parameter sets, deeper parameter sets tended to have more robust performance in differential split sample testing. Here ‘deeper’ refers to parameter sets with higher data depth as defined by Tukey (1975). Data depth was used in the present study when quantifying the performance of DREAM-ABC behavioral ensembles for the purpose of comparison of calibration methods. We calculated the data depth for every parameter set in each DREAM-ABC ensemble, using the method of Rousseeuw and Struyf (1998). For each ensemble, the results from its deepest set were adopted to represent the ensemble in comparisons with other methods. As explained in the introduction, identification of the ‘deepest’ set in an ensemble considers the position in parameter space of every other member in the ensemble, making it a better candidate to ‘represent’ the ensemble than the median or the numerically optimum parameter set.

However, it is interesting to analyse whether these deeper parameter sets provide better performance in split sample testing, for the present study catchments, model structures, and calibration methods. Figure D.6 examines the performance in split sample tests of these deepest sets, compared with the median performance from the parent ensemble. This figure shows two main messages: (i) in cases where models perform well in split sample testing (for example, $KGE_{dry} > 0.6$ as per the grey box in Figure D.6), the deepest parameter sets provide a higher $KGE_{dry}$ than
Figure D.4: Hybrid method results for the $C_v$ annual signature for GR4J. Interpretation is similar to Figure 6.2 in Chapter 6. In the square brackets, the first number is the KGE_{dry} obtained by the reference method, and the second number is that obtained by the hybrid method. Catchments selected for inclusion in Chapter 5 are marked with a grey box.
Figure D.5: Hybrid method results for the High Flow Discharge signature for SIMHYD. Interpretation is similar to Figure 6.2 in Chapter 6. In the square brackets, the first number is the KGE\text{dry} obtained by the reference method, and the second number is that obtained by the hybrid method. Catchments selected for inclusion in Chapter 5 are marked with a grey box.
Figure D.6: Performance of sets of high data depth in split sample testing, for every case study shown in Figure 6.5. KGE in the dry (evaluation) period (KGE\textsubscript{dry}) for the deepest parameter set is plotted on the y-axis. The x-axis plots the median KGE\textsubscript{dry} across the parent ensemble. The results show that deeper parameter sets generally have more robust performance in split sample testing if: (i) the model structure is GR4J, SIMHYD or GR4JMOD; and/or (ii) model performance in the dry period is generally good (e.g., KGE\textsubscript{dry} > 0.6 as per the grey box). Values less than zero exist but are not shown.

the median of their parent ensemble; and (ii) in cases where KGE\textsubscript{dry} < 0.6, the relative performance of the deepest sets is model specific: the deepest sets work better only for GR4J, GR4JMOD and SIMHYD. For IHACRES and SACRAMENTO, the deepest set provides no advantage relative to the rest of the ensemble unless (i) holds.
D.7 Repeat of key results using alternative evaluation metrics

Although Chapter 6 mainly relies on a single metric, the Kling Gupta Efficiency (KGE, Gupta et al., 2009) to express model performance, no single metric gives a complete picture of model performance. Therefore, in this section we re-compile results based on different performance metrics to the KGE.

In total, six metrics are shown, taken from the recommended lists of Thirel et al. (2015b) for performance metrics suitable for use in a changing climate:

- KGE (shown in Chapter 6, Figure 6.1) - note that we use the original formulation of Gupta et al. (2009);

- Nash Sutcliffe Efficiency (NSE) - Nash and Sutcliffe (1970)

- Bias, expressed as a ratio of simulated and observed means

- Linear correlation (Pearson Correlation)

- Relative variability (ratio of simulated and observed variability)

- NSE low flow, or NSE$_{LF}$. For more information, refer to Chapter 5, Table 5.1 and Pushpalatha et al. (2012).

For brevity and since the main recommendations of this study concern objective functions (specifically the Revised Index of Agreement and the Split KGE), we provide these plots only for Figure 6.1 (objective functions) and not for Figure 6.33 (hybrid methods) or Figure 6.5 (DREAM-ABC).
Figure D.7: Repeat of Figure 6.1 from Chapter 6, using the Nash-Sutcliffe Efficiency (NSE) as the reference metric.
Figure D.8: Repeat of Figure 6.1 from Chapter 6, using the bias as the reference metric. Here bias is expressed as a ratio of means ($\mu_{\text{sim}} / \mu_{\text{obs}}$). Values higher than 3.0 exist but are not shown.
Figure D.9: Repeat of Figure 6.1 from Chapter 6, using the Linear Correlation (Pearson’s Correlation Coefficient) as the reference metric.
Figure D.10: Repeat of Figure 6.1 from Chapter 6, using relative variability as the reference metric. Here relative variability is expressed as $\sigma_{sim} / \sigma_{obs}$ (Thirel et al., 2015b). Values greater than 2 exist but are not shown.
Figure D.11: Repeat of Figure 6.1 from Chapter 6, using the NSE$_{LF}$ as the reference metric. For more information on this metric, refer to Chapter 5, Table 5.1 and Pushpalatha et al. (2012).
Appendix E

Supplementary material for
Chapter 7

E.1 Introduction

This appendix contains the following:

1. Notes on application of Step 4a for case study catchment (Section E.2);

2. Additional methods of analysis of model structure for Step 4b (Section E.3);

3. General notes on application of Step 5b, non case-specific (Section E.4);

4. Example application of framework to multiple catchments (Section E.5); and

5. Selected plots and tables:

   (a) Table of parameters for IHACRES-A and IHACRES-B including thresholds (Table E.1);

   (b) Table of possible points of investigation for Step 4a (non case-specific) (Table E.2); and

   (c) Simulation dynamics: additional plots comparing IHACRES-A and IHACRES for the case study catchment (Figure E.4);
E.2 Notes on application of Step 4a for case study catchment

Step 4a calls for an analysis of change in data errors, human impacts or natural processes. Notable temporal changes in data for the Harvey River catchment include the decommissioning of four rain gauges in the year 2000. This left no gauges in the catchment, implying that the gridded data that provides the basis for the rainfall in this study (cf. Section 7.4.1) was based on a greater degree of spatial interpolation after 2000 than before. However, double mass curve analysis does not reveal a breakpoint at this time (Figure E.1), possibly because the outside gauges are still relatively nearby (<3km from the catchment boundary) and because rainfall spatial patterns are relatively uniform (Jones et al., 2009), being dominated by large scale frontal systems. Wind data from (Donohue et al., 2010, p27) indicates an upward trend in wind speed in the region, so underestimation of evaporative demand in the latter years of the timeseries by the Morton Wet Environment method is a possible hypothesis as a cause for the model structural failure. However, it is unlikely that all of the problem could be due to this factor, given the large discrepancies between simulated and observed flow (eg. bias of 43%). Analysis of flow ratings and uncertainty was conducted by McMahon, T. and Peel, M., Uncertainty in stage-discharge rating curves: Application to Australian Hydrologic Reference Stations data, submitted to Hydrologic Sciences Journal, 12th June2017. Their analysis revealed that station 613002 has high data quality and low uncertainty relative to most other stations in their dataset (171 stations), with minimal extrapolation and no significant changes to the rating curve with time.

Considering human impacts, the catchment is within Perth’s water supply area and has historically been clear of water impoundments, pumping, and treatment plants. It has remained forested throughout the period of historic observations. Thus, it is considered unlikely that the model structural failure is due to human impacts. With respect to vegetation, to our knowledge no studies of changes in time in vegetation characteristics or dynamics have been conducted in the catchment. Fires have occurred historically in the catchment, as outlined by Wittkuhn et al. (2010).
Figure E.1: Double mass curves comparing catchment average rainfall for the case study catchment 613002 (x axis) with catchment average rainfall for the four closest catchments (y axis) from Turner (2012). The blue lines depict cumulative rainfall over 1900 to 2014, compared to the 1 to 1 line (dashed). All axes vary between 0 and 150000 mm. Based on daily catchment average rainfall from Jones et al. (2009). There are no breaks of slope apart from 610008 at the very beginning of the timeseries. This indicates that the removal of numerous rain gauges from within 613002 (cf. Section 7.4.9) in the year 2000 did not cause any systematic bias changes with time, relative to nearby catchments.

The catchment is regularly subject to planned burns intended to reduce combustible material without killing the Jarrah forest. Since 1960, the only significant wildfire occurred during the DSST evaluation period (December 2009) and affected over half of the catchment. However, it is unlikely that the 2009 bushfire explains the over-estimation of flow by the model in the evaluation period, for two reasons: (i) the December 2009 event is very close to the end of the evaluation period (December 2011); and (ii) increases, not decreases, in runoff are expected initially following the bushfire (as the transpiration demand is temporarily lessened prior to re-sprouting).

When considering changes in catchment behaviour, environmental measurements relevant to rainfall-runoff model state variables may help to characterise qual-
itative differences between simulations and reality. Examples include groundwater levels, soil moisture measurements (in situ or remotely sensed), and water chemistry or isotopic data. For the present case study, lack of data are the main barrier to characterisation of natural processes. Two groundwater bores in the public database are within the catchment boundaries, but the associated records no longer exist or are not publically available (cf. Hughes et al., 2012); likewise for thirteen nearby bores to the north and east. A further six bores have records only for 1988-1992 and are thus unhelpful in characterising long term trends. However, studies of catchments in the region indicate a correlation between decreasing runoff ratios and declining groundwater levels (Hughes et al., 2012; Kinal and Stoneman, 2012).

In summary, the answer to the key question (are there any temporal changes in data errors, human impacts or natural processes that could help to explain the Pareto failure?) is that there is no evidence of temporal changes in rainfall or runoff data errors, but trends in wind may have caused changes in evaporative demand that are not characterised in the adopted PET data. Being an undeveloped catchment, it is unlikely that changes in human practices are impacting on flow, and changes in natural processes (eg., dynamics of vegetation and groundwater) are difficult to assess due to lack of data.

E.3 Additional methods of analysis of model structure for Step 4b

Relevant to Section 7.4.10, this text explains two methods used to examine whether parameters change with time when fitted to sub-periods of calibration data. As explained in Chapter 7.4.9, a separate analysis is done over longer (years-decades) decades, and shorter (weeks-months) timescales (the latter to examine seasonal transitions from dry to wet).

To examine long timescales, we use the two contrasting periods defined in Section 7.4.5, identifying an ensemble of ‘near-optimal’ sets for 1970-2004 (wetter), then separately for 2005-2011 (drier). Near-optimal is defined as the top 0.1% by
KGE value, in a random (Latin Hypercube) sample (N = 3ÅU106). This provides 3000 parameter sets for each period, and histograms of values for each IHACRES-
A parameter are derived, along with Kolmogorov-Smirnov D values as a relative
measure of difference between the distributions (note these values are not used for
significance testing). Although the use of histograms differs from previous studies
(eg. Merz et al., 2011), the underlying logic is similar.

The results (Figure E.2) indicate that the c parameter changes most between
the two periods, with lower values in the red histogram than in the blue. Parameters
Tw and f have low identifiability, possibly indicative of parameter interaction. The
routing parameters are relatively stable in time, although Tf decreases slightly.

For short timescales, Dynamic Identifiability Analysis (Wagener, 2003, DYNIA) is implemented using the Sensitivity Analysis For Everyone (SAFE) toolkit
(Pianosi et al., 2015). DYNIA analyses the dynamics over rolling time-windows of
a fixed length (often less than a month). A random ensemble of parameter sets is
analysed, and for each time window, a fixed percentage of best-performing parame-
ter sets are identified. Distributions of parameters within this percentage are then
derived, and DYNIA displays temporal changes in these distributions. SAFE toolkit
default settings are used here (window length = 21 days; % retained = 10; selection
basis: least squares), with a random sample of 20,000 sets.

DYNIA results show that the c parameter changes the most with time, and
is highly identifiable. For illustration, Figure E.2 shows DYNIA results for the year
1999, for the parameter c and (for contrast) a non-identifiable parameter, f. Para-
meter f has a time-invariant uniform distribution, which (as above) may indicate
parameter interaction. Parameter c has lower values in the dry season (c<0.0005)
than during the wet season (0.0005>c>0.0015). Values during the wet period fluctu-
ate with time, indicating possible interaction with routing components and timing
errors. The transition between the dry and wet, referred to here as the ‘wetting up’
period, includes times (marked by the grey box) where c values are lower relative to
the wet season. Calibrated models where c is chosen to match the wet season as a
whole would overestimate flow in the wetting up period.
Figure E.2: Selected analysis tool for model structural evaluation. Part 1: Shift in parameters over different time periods, for free parameters in IHACRES-A. Histograms show the distribution of parameter values for ‘near optimal’ (see text for definition) parameter sets over two contrasting periods. Results show the c parameter changes most. Part 2: DYNIA outputs for a sensitive parameter, c, and non-sensitive parameter, f, over a single year (1999). In the case of c, allowing the parameter value to change to fit brief windows of time results in high temporal variability.

E.4 Notes - general application of Step 5b (not case-study-specific)

Step 5b involves the addition of one or more alternative data types to the calibration process in the hope that this will improve performance in DSST evaluation. However, Step 5b may be undertaken for other reasons such as increasing the realism of the model (Clark et al., 2011). Many prior studies demonstrate calibration using multiple data types, and this section summarises available options.

Multi-response calibration studies are common in the literature, often driven by the need to parameterise high dimensional physically-based or distributed mod-
els (Vaché and McDonnell, 2006; Clark et al., 2011; Rakovec et al., 2016). The approach depends both on data availability and the degree to which model fluxes or state variables correspond with real-life ‘observables’. Data of various types can be incorporated, including groundwater levels (e.g. Seibert et al., 1997; Lamb et al., 1998; Seibert, 2000; Beldring, 2002), soil moisture measurements (Kolma et al., 1995; Wooldridge et al., 2003; Brocca et al., 2010), water quality information (Mroczkowski et al., 1997; Kuczera and Mroczkowski, 1998; Muleta and Nicklow, 2005; Dean et al., 2009), snow cover data (Parajka et al., 2005) and remotely-sensed information such as Leaf Area Index, evaporation, or water storage data (Hughes et al., 2013; Tesemma et al., 2015; Winsemius et al., 2009; Rakovec et al., 2016). Although not all studies report reduced parameter uncertainty (Kuczera and Mroczkowski, 1998), the “scrutiny of multivariate data ... not only exposes major model deficiencies, but may be indispensable for improving model realism” (Clark and Kavetski, 2010b). Thus, as noted, multi-response data may be relevant at various stages of this framework, not just calibration. Bayesian approaches can quantify changes in parameter identifiability, predictive uncertainty, and/or rejection of model hypotheses, resulting from additional data (Beven and Binley, 1992; Mroczkowski et al., 1997), but other multi-objective calibration paradigms may also be useful, eg. Pareto approaches can define trade-offs between matching different data types (Efstathiadis and Koutsogiannis, 2010).

In most real-life applications, continuous monitoring of the above variables will be unavailable or patchy, and model calibration using less data-intensive approaches may be required. ‘Soft data’ was initially defined by Seibert and McDonnell (2002) as “qualitative knowledge... that cannot be used directly as exact numbers”, and other studies have used the term to describe all manner of non-continuous numerical data and fuzzy expert knowledge. Working in the heavily instrumented Mai Mai catchment in New Zealand, Seibert and McDonnell (2002) incorporated two soft data types: qualitative judgment of ‘reasonability’ of parameter values; and fuzzy rules matching simulated groundwater behaviour and ‘new water’ fraction to qualitative expectations of experimental hydrologists. Winsemius et al. (2009) included recession curve signatures, spectral properties and intercomparison between daily- and
monthly-timestep model outputs in their definition of soft data. Gharari et al. (2014) included logical parameter rules (eg. that wetlands should have shallower unsaturated zones than hillslopes) and snapshots of NDVI values based on remote sensing. Following a controlled experiment that recorded how ten modellers responded to a data-poor example application, Holländer et al. (2014) emphasised the importance of subjective knowledge, including field visits, interactions with experts, and a-priori parameter estimation. They concluded that “process understanding can be as efficient as adding data for improving parameters”. All of these studies reported either reduced predictive uncertainty or improved identifiability by augmenting traditional calibration data with soft data/qualitative knowledge. Although the availability and type of soft data varies with context, these examples demonstrate the diversity of soft data approaches. In summary, there are many ways in which additional data may be incorporated in model calibration.

E.5 Example application of framework to multiple catchments

One benefit of the framework is the ability to upscale from one catchment to many, allowing potential structural changes to be assessed for generality across multiple locations. As an example, in Figure E.3 we repeat the analysis of Figure 7.5 for the set of 86 catchments used in Chapter 4. Instead of plotting one Pareto curve per catchment, a single summary statistic is extracted from each Pareto curve (closest distance to the ‘perfect point’; see Figure 7.5a) and Figure E.3 shows the distribution of values obtained across catchments, for IHACRES-A and IHACRES-B separately. The IHACRES-B Pareto curves are much closer to the ‘perfect’ point, on average, than the IHACRES-A curves. Thus, in general, the extra parameters of (Ye et al., 1997, cf. Section 7.4.2) allow an increase in robustness of simulations, across the full geographic range of the source dataset (in this case southern Australia). The analysis can be extended to the hydrologic signatures also. Figure E.3b shows that error in mean and median tends to be less climate sensitive for IHACRES-B than
IHACRES-A - results consistent with (albeit more muted than) the Harvey River catchment. The slow flow and recession signatures also show improvement across the full set of catchments, despite being worsened in Harvey River (Figure 7.5). Given the importance of large sample hydrology studies to develop generalisable conclusions (Gupta et al., 2014), the ability of this framework to upscale across many catchments is a clear advantage.

E.6 Selected tables and figures
Figure E.4: Comparison of simulation dynamics for IHACRES-A and IHACRES-B, shown alongside observed variables. (a) and (b) show the proportion of rainfall converted to ‘effective’ rainfall (runoff), which is calculated for each timestep as a function of model state*. For comparison, (c) and (d) show the observed precipitation and runoff, respectively. Simulation dynamics for IHACRES-A vary little from year to year. In contrast, IHACRES-B dynamics show greater sensitivity to small changes in rainfall, which is qualitatively similar behavior to observed runoff. Simulations are based on parameter sets enlarged in Figure 7.4. *Specifically, (a) shows the catchment wetness index, sk (using the terminology of (Jakeman and Hornberger, 1993) and (b) shows [(sk - l)p] which reflects the extra parameters added by (Ye et al., 1997).
### Table E.1: Parameter list for IHACRES-A and IHACRES-B

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Min</th>
<th>Max</th>
<th>IHAC-A</th>
<th>IHAC-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tw (days)</td>
<td>drying rate at reference PET$^*$</td>
<td>1</td>
<td>200</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>F (-)</td>
<td>Drying rate exponent</td>
<td>0</td>
<td>4</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>c (mm-1)</td>
<td>Scale parameter$^+$</td>
<td>0.00001</td>
<td>1</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Tq (days)</td>
<td>Time constant for quick store$^*$</td>
<td>0.00001</td>
<td>200</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Ts (days)</td>
<td>Time constant for slow store$^*$</td>
<td>0.00001</td>
<td>200</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Vs (-)</td>
<td>Split between quick and slow</td>
<td>0</td>
<td>1</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>p (-)</td>
<td>Runoff production exponent</td>
<td>1</td>
<td>5</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>l (mm)</td>
<td>Wetness threshold for runoff</td>
<td>0</td>
<td>200</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

*$^*$ Tw, Tq & Ts are defined as time constants, i.e. number of time steps to reduce to $1/e$ (approx 37%)

*$^+$ c can be interpreted as similar to the inverse of the soil moisture storage capacity

**reference PET set to 0 mm/d**
Table E.2: Possible points of investigation for Step 4a (Analysis of change in data errors, human impacts or natural processes). Note: not specific to any case study.

<table>
<thead>
<tr>
<th>Data errors and related issues</th>
<th>Human impacts</th>
<th>Natural processes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rain gauge errors</strong>: trends due to obstructions (eg. growth of trees) or temporal changes in wind (Nešpor and Sekruk, 1999; Donohue et al., 2010).</td>
<td><strong>Water impoundments</strong>: Variable temporal impacts due to reservoirs or small on-farm impoundments (farm dams/farm ponds) that are unaccounted for by the model (Schreider et al., 2002; Nathan and Lowe, 2012; Fowler et al., 2016). Temporal trends in construction of these impoundments.</td>
<td><strong>Geomorphologic change including erosion, bedform changes, landslides or stream capture</strong> (Dunne, 1980; Bishop, 1995; Øygarden, 2003).</td>
</tr>
<tr>
<td><strong>PET assumptions</strong>, eg. some formulations do not consider wind speed, which may trend with time (Donohue et al., 2010; McMahon et al., 2013).</td>
<td><strong>Surface water extractions</strong>: temporal changes in pumping of water out of the stream for local use or inter-basin transfer (Ferguson and Maxwell, 2012).</td>
<td><strong>Changes in vegetation dynamics due to different climatic conditions</strong>, eg. temporal trends in species mix (Troch et al., 2009), growing season length (Merz et al., 2011) or water use efficiency due to different climatic conditions (Huxman et al., 2004; Troch et al., 2009).</td>
</tr>
<tr>
<td><strong>PET source data changes</strong>: PET relies on other data that may be subject to changes in technology or practice; eg. solar radiation may currently be directly measured but previously inferred from sunshine hours.</td>
<td><strong>Treatment plants</strong>: changes in behaviour of treatment plants removing or adding water (Carey and Migliaccio, 2009).</td>
<td><strong>Fire</strong>: If severe, wildfire may cause temporary increases in flow followed by decreases as juvenile trees establish and use more water than mature trees (Kuczera, 1987; Cornish and Vertessy, 2001).</td>
</tr>
<tr>
<td><strong>Gridded products source data changes</strong>: Bias of gridded products (rain, temperature) may change with time as gauges are commissioned or decommissioned (Jones et al., 2009; Stokstad, 1999).</td>
<td><strong>Groundwater extraction causing temporal trends in groundwater-surface water interactions</strong> (Barlow and Leake, 2012).</td>
<td><strong>Catchment storage and ‘memory’</strong>: Storage and release of water is central to hydrologic response (Mcnamara et al., 2011). Storage dynamics may provide a mechanism for the effect of climatic anomalies to persist over multiple subsequent years (Yang et al., 2017), or accumulate with time if such conditions persist (Hughes et al., 2012; Kinal and Stoneman, 2012; Saft et al., 2016b). Most hydrological models do not have this property.</td>
</tr>
<tr>
<td><strong>Flow gauge infrastructure may develop bias due to erosion or other forms of degradation</strong> (McMillan et al., 2010). Gauges may be moved, causing a step change in the error structure of the rating curve.</td>
<td><strong>Landuse-induced transpiration changes</strong> due to deforestation, changes in irrigation practice, changes in agricultural practice, or changes in plantations (Cornish and Vertessy, 2001; Neal et al., 2001; Breuer et al., 2009; Ferguson and Maxwell, 2012; McIntyre et al., 2013).</td>
<td><strong>Change to a different stable state</strong>: eg. Peterson et al. (2009) demonstrated a catchment model with multiple stable states due to feedbacks between vegetation water use and groundwater level of a saline aquifer.</td>
</tr>
<tr>
<td><strong>Flow rating retrospectivity</strong>: issues may arise if data from new ratings are not retrospectively applied as appropriate.</td>
<td><strong>Urbanisation causing changes to hydrologic response</strong> (Leopold et al., 2005; Paul and Meyer, 2001; ODriscol et al., 2010).</td>
<td></td>
</tr>
</tbody>
</table>
References


REFERENCES


REFERENCES


Bergström, s., B. Carlsson, M. Gardelin, G. Lindström, A. Petterson, and M. Rummukainen (2001), Climate change impacts on runoff in Sweden - Assessments by global climate models, dynamical downscaling and hydrological modelling, Climate Research, 16(2), 101–112, doi: 10.3354/cr016101. 8, 100, 144, 184


REFERENCES


REFERENCES


Bowling, D. R., E. S. Schulze, and S. J. Hall (2017), Revisiting streamside trees that do not use stream water: can the two water worlds hypothesis and snowpack isotopic effects explain a missing water source?, Ecohydrology, 10(1), e1771-n/a, doi: 10.1002/eco.1771, e1771 ECO-16-0104.R1. 40
REFERENCES


Broderick, C., T. Matthews, R. L. Wilby, S. Bastola, and C. Murphy (2016), Transferability of hydrological models and ensemble averaging methods between


Castiglioni, S., L. Lombardi, E. Toth, A. Castellarin, and A. Montanari (2010),
REFERENCES


Chai, T., and R. R. Draxler (2014), Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature, *Geoscientific Model Development*, 7(3), 1247–1250, doi: 10.5194/gmd-7-1247-2014. 135, 141


REFERENCES


Cook, B. I., K. J. Anchukaitis, R. Touchan, D. M. Meko, and E. R. Cook (2016), Spatiotemporal drought variability in the Mediterranean over the last 900


REFERENCES


REFERENCES


REFERENCES


Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research, 28*(4), 1015–1031. 21, 22, 72, 107, 137, 146


Eberhart, R. C., Y. Shi, and J. Kennedy (2001), *Swarm intelligence*, Elsevier. 62


REFERENCES


Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, Water Resources Research, 32(7), 2161–2173, doi: 10.1029/96WR03723. 90, 102, 151, 188, 212


Fu, G., F. H. S. Chiew, S. P. Charles, and F. Mpelasoka (2011), Assessing precipitation elasticity of streamflow based on the strength of the precipitation-streamflow relationship, in 19th International Congress on Modelling and Simulation (MODSIM), December, pp. 3567 – 3572, Perth, Australia. 9


Gao, H., M. Hrachowitz, S. J. Schymanski, F. Fenicia, N. Srwongsitanon, and H. H. G. Savenije (2014), Climate controls how ecosystems size the root zone
storage capacity at catchment scale, Geophysical Research Letters, 41(22), 7916-7923, doi: 10.1002/2014GL061668. 39, 186


Hansen, N., S. D. Müller, and P. Koumoutsakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evolutionary computation, 11*(1), 1–18, doi: 10.1162/106365603321828970. 22, 72, 138


REFERENCES


Hutton, C., T. Wagener, J. Freer, D. Han, C. Duffy, and B. Arheimer (2016), Most
computational hydrology is not reproducible, so is it really science?, *Water Resources Research*, 52(10), 7548–7555, doi: 10.1002/2016WR019285. 21, 25


REFERENCES


Kiern, A. S. (2004), Multi-decadal variability of drought risk, eastern Australia,
*Hydrological Processes*, 18(11), 2039-2050, doi: 10.1002/hyp.1460. 5, 6, 12

Kim, H. S., and S. Lee (2014), Assessment of a seasonal calibration technique using
multiple objectives in rainfall-runoff analysis, *Hydrological Processes*, 28(4), 2159-
2173, doi: 10.1002/hyp.9785. 27, 61

Kinal, J., and G. Stoneman (2012), Disconnection of groundwater from surface water
causes a fundamental change in hydrology in a forested catchment in south-western

Kirchner, J., X. Feng, and C. Neal (2000), Fractal stream chemistry and its impli-
cations for contaminant transport in catchments, *Nature*, 403(6769), 524-7, doi:
10.1038/3500537. 34

Kirchner, J. W. (2003), A double paradox in catchment hydrology and geochemistry,
*Hydrological Processes*, 17(4), 871-874, doi: 10.1002/hyp.5108. 34

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking
measurements, analyses, and models to advance the science of hydrology, *Water

Kirchner, J. W. (2009), Catchments as simple dynamical systems: Catchment char-
acterization, rainfall-runoff modeling, and doing hydrology backward, *Water Res-

Kirchner, J. W., X. Feng, and C. Neal (2001), Catchment-scale advection and dis-
persion as a mechanism for fractal scaling in stream tracer concentrations, *Journal
of Hydrology*, 254(1-4), 82-101, doi: 10.1016/S0022-1694(01)00487-5. 34

Klemes, V. (1983), Conceptualization and scale in hydrology, *Journal of Hydrology,
65*(1-3), 1-23, doi: 10.1016/0022-1694(83)90208-1. 34, 96

Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrolog-
56, 58, 96, 102, 103, 158, 185, 192, 216, 217


Koutsoyiannis, D., G. Blöschl, A. Bárcroixy, C. Cudennec, D. Hughes, A. Montanari,


Kuczera, G., B. Renard, M. Thyer, and D. Kavetski (2010), There are no hydrological monsters, just models and observations with large uncertainties!, *Hydrological Sciences Journal, 55*(6), 980–991, doi: 10.1080/02626667.2010.504677. 24


Lamb, R., and K. J. Beven (1997), Using interactive recession curve analysis to


Mann, H. B., and D. R. Whitney (1947), On a test of whether one of two random variables is stochastically larger than the other, *The annals of mathematical statistics*, pp. 50–60. 85


REFERENCES


Michelson, D. B. (2004), Systematic correction of precipitation gauge observations


REFERENCES


Øygarden, L. (2003), Rill and gully development during an extreme winter runoff event in norway, Catena, 50(2), 217–242. 314


Park Williams, A., and C. Funk (2011), A westward extension of the warm pool leads to a westward extension of the walker circulation, drying eastern africa, Climate Dynamics, 37(11), 2417–2435, doi: 10.1007/s00382-010-0984-y. 1, 6


REFERENCES


Potter, N. J., and F. H. S. Chiew (2009), Statistical characterisation and attribution of recent rainfall and runoff in the Murray-Darling Basin, in 18th World IMACS / MODSIM Congress, July, pp. 2812 – 2818, Cairns, Australia. 13, 38


Power, S., T. Casey, C. Folland, A. Colman, and V. Mehta (1999), Inter-decadal modulation of the impact of ENSO on Australia, Climate Dynamics, 15(5), 319–324, doi: 10.1007/s003820050284. 5, 6, 12


REFERENCES


Refsøgaard, J. C., H. Madsen, V. Andréassian, K. Arnbjerg-Nielsen, T. A. Davidson,
REFERENCES


REFERENCES


REFERENCES


Schaeflri, B., and H. V. Gupta (2007), Do nasal values have value?, *Hydrological Processes*, 21(15), 2075–2080, doi: 10.1002/hyp.6825. 177


REFERENCES


Sinclair Knight Merz, CSIRO, and Bureau of Regional Sciences (2010), Surface and/or groundwater interception activities: initial estimates, *Report, online*, National Water Commission Canberra, A.C.T. 43
Siriwardena, L., A. Western, B. Anderson, and F. Chiew (2008), Prediction of
Within-day Rainfall Intensity Distributions from Daily Rainfall and Meteorologi-
cal Data in Australia, in Hydrology and Water Resources Symposium (Water Down
Under), pp. 2241–2252. 10

Sivapalan, M., G. Blöschl, L. Zhang, and R. Vertessy (2003), Downward ap-
proach to hydrological prediction, Hydrological Processes, 17(11), 2101–2111, doi:
10.1002/hyp.1425. 35

Smith, M., D. Laurine, V. Koren, S. Reed, and Z. Zhang (2003), Hydrologic model
calibration in the National Weather Service, in Advances in the Calibration of
Watershed Models, American Geophysical Union. 20, 21, 145

Smith, P. J., K. J. Beven, and J. A. Tawn (2008), Detection of structural inade-
quacy in process-based hydrological models: A particle-filtering approach, Water
Resources Research, 44(1), doi: 10.1029/2006WR005205. 27

Smith, T., L. Marshall, B. McGlynn, and K. Jencso (2013), Using field data to inform
and evaluate a new model of catchment hydrologic connectivity, Water Resources
Research, 49(10), 6834–6846, doi: 10.1002/wrcr.20546. 35

Smith, T., L. Marshall, and A. Sharma (2015), Modeling residual hydrologic
errors with bayesian inference, Journal of Hydrology, 528, 29 – 37, doi:
https://doi.org/10.1016/j.jhydrol.2015.05.051. 102

Sobol, I. M. (2001), Global sensitivity indices for nonlinear mathematical models and
their Monte Carlo estimates, Mathematics and Computers in Simulation, 55(1-3),
271–280, doi: 10.1016/S0378-4754(00)00270-6. 22, 37, 205
REFERENCES


Tesemma, Z., Y. Wei, M. Peel, and A. Western (2015), The effect of year-to-year variability of leaf area index on variable infiltration capacity model performance and simulation of runoff, Advances in Water Resources, 83, 310–322. 309
REFERENCES


Trenberth, K. E. (2011), Changes in precipitation with climate change, *Climate Research, 47*(1/2), 123-138. 1, 6, 14, 226


Tukey, J. (1975), Mathematics and the picturing of data, in *Proceedings of the international congress of mathematicians*, pp. 523-532. 31, 103, 147, 157, 292


Turner, M. (2012), Hydrologic Reference Station Selection Guidelines. 65, 197, 267, 305


Global Change Research Program, 1 online resource (188 p.) : pp., Cambridge University Press New York. 1, 6


van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration


REFERENCES


REFERENCES


REFERENCES


Wittkuhn, R. S., T. Hamilton, et al. (2010), Using fire history data to map temporal sequences of fire return intervals and seasons, Fire Ecology, 6(2), 97-114. 304


World Bank (2016), High and Dry: Climate Change, Water, and the Economy, Working papers, eSocialSciences. 1, 104


Yang, Y., T. R. McVicar, R. J. Donohue, Y. Zhang, M. L. Roderick, F. H. Chiew, L. Zhang, and J. Zhang (2017), Lags in hydrologic recovery following an extreme


Author/s:
Fowler, Keirnan

Title:
Towards improved rainfall-runoff modelling in changing climatic conditions

Date:
2017

Persistent Link:
http://hdl.handle.net/11343/208776

File Description:
Towards improved rainfall-runoff modelling in changing climatic conditions

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.