Reliable long-range ensemble streamflow forecasts by combining dynamical climate forecasts, a conceptual runoff model and a staged error model

James C. Bennett[1], QJ Wang[1], Ming Li[2], David Robertson[1], Andrew Schepen[3]

[1]CSIRO Land & Water, Clayton, Victoria, Australia

[2]Data61, Floreat, Western Australia, Australia.

[3]CSIRO Land & Water, Dutton Park, Queensland, Australia

Corresponding author: James Bennett (james.bennett@csiro.au)

Key Points:

- A new method to produce statistically reliable ensemble forecasts of streamflow to 12 months ahead
- Forecasts are designed to be skillful alternatives to stochastic scenarios
- Forecasts of accumulated volumes (e.g. 6-month total flow) can be skillful to long time horizons

Abstract

We present a new streamflow forecasting system called forecast guided stochastic scenarios (FoGSS). FoGSS makes use of ensemble seasonal precipitation forecasts from a coupled ocean-atmosphere general circulation model (CGCM). The CGCM forecasts are post-processed with the method of calibration, bridging and merging (CBaM) to produce ensemble precipitation forecasts over river catchments. CBaM corrects biases and removes noise from the CGCM forecasts, and produces highly reliable ensemble precipitation forecasts. The post-processed CGCM forecasts are used to force the Wapaba monthly rainfall-runoff model. Uncertainty in the hydrological modelling is accounted for with a 3-stage error model. Stage 1 applies the log-sinh transformation to normalize residuals and homogenize their variance; Stage 2 applies a conditional bias-correction to correct biases and help remove negative forecast skill; Stage 3 applies an autoregressive model to improve forecast accuracy at short lead-times and propagate uncertainty through the forecast. FoGSS generates ensemble forecasts in the form of time series for the coming 12-months.

In a case study of two catchments, FoGSS produces reliable forecasts at all lead-times. Forecast skill with respect to climatology is evident to lead-times of about 3 months. At longer lead-times, forecast skill approximates that of climatology forecasts; that is, forecasts become like stochastic scenarios. Because forecast skill is virtually never negative at long lead-times, forecasts of accumulated volumes can be skillful. Forecasts of accumulated 12-month streamflow volumes are significantly skillful in several instances, and ensembles of accumulated volumes are reliable. We conclude that FoGSS forecasts could be highly useful to water managers.

1. INTRODUCTION

Many water agencies rely on simple resampling of historical streamflow (stochastic scenarios) to plan operations of water storages over timescales of up to 12 months. This approach is practical: stochastic scenarios are simple to generate to long times horizons; they are in the form of time series, which is often a requirement for water planning models; they are inherently realistic - i.e., they have correct autocorrelation and seasonal cycles and they are unbiased; it is easy to generate an ensemble of scenarios to give a range of plausible possibilities. These properties allow water managers to run stochastic scenarios directly through quantitative planning models. The obvious disadvantage of stochastic scenarios is that they ignore the current states of the catchment and the climate, which can be used to produce informative streamflow forecasts. As several studies have demonstrated, information from streamflow forecasts can allow water managers to allocate water resources more profitably [*Crochemore et al.*, 2015; *Ramos et al.*, 2013].

Recent years have seen the rise of sophisticated ensemble seasonal hydrological forecasting systems and services, encouraged by research communities such as the Hydrological Ensemble Prediction EXperiment (HEPEX, www.hepex.org). These systems are capable of producing skillful forecasts, either with statistical techniques [e.g. *Bracken et al.*, 2010; *Grantz et al.*, 2005; *Regonda et al.*, 2006; *Wang et al.*, 2009], a combination of statistical methods and conceptual models [e.g. *Bradley et al.*, 2015; *Najafi et al.*, 2012; *Werner et al.*, 2004], or through the use of dynamical and/or conceptual models to forecast climate and land surface processes [see review by *Yuan et al.*, 2015a]. However, there are usually some aspects of these forecasting methods that lack the convenience of stochastic scenarios for water managers: they may forecast total volumes over seasons rather than producing forecasts that are broken down into time series; they often do not produce forecasts for the long time horizons (e.g. 12 months) that water managers require; the spread of the ensemble may be too wide (underconfident) or, more typically, too narrow (overconfident); the forecasts may be less informative than climatology for some months or lead-times [i.e., forecasts may have negative skill; see, e.g., *Olsson et al.*, 2016]. In our experience, any of these can be a serious barrier for water managers who wish to use forecasts in quantitative planning models.

The spread of the ensemble is of particular note. Overconfidence in forecast ensembles can lead to rash decisions, while underconfidence can result in decisions that are overly cautious [*Arnal et al.*, 2016]. We use the term *reliability* to describe the aptness of the ensemble spread, following long-standing convention [e.g., *Murphy*, 1977]. A forecast ensemble that is correctly distributed (neither over- nor underconfident) is reliable.

In this paper, we present a new seasonal streamflow forecasting system, called forecast guided stochastic scenarios (FoGSS). FoGSS aims to combine the convenience of stochastic scenarios with the skill of a modern forecasting system. The first task in establishing FoGSS is to generate forecasts that are as skillful as possible. Predictability of seasonal streamflow is commonly ascribed to two sources: i) the state of the climate (atmosphere and oceans) and ii) the state of the land surface, including soil moisture, ground water, snow and ice amounts and their fluxes [e.g. *Koster et al.*, 2010; *Mahanama et al.*, 2012; *Yuan et al.*, 2015a]. The state of the climate is necessary to predict precipitation; the state of the land surface measures water already in the catchment, but which has not yet entered rivers.

Coupled ocean-atmosphere general circulation models (CGCMs) are widely considered the state-of-the-art approach for predicting seasonal climate. CGCMs assimilate recent observations of the oceans and the atmosphere, and use physical and empirical equations to forecast climate responses to initial conditions. CGCMs have been shown to outperform statistical methods in forecasting sea surface temperature to long lead-times [*Barnston et al.*, 2012] and can show skill in predicting precipitation to long lead-times (>6 months) in certain locations and seasons [e.g. *Hawthorne et al.*, 2013]. Many weather agencies base seasonal climate outlooks on CGCM forecasts, and a number of weather agencies (e.g. the UK Met Office, the European Centre for Medium range Weather Forecasting, the US National Oceans and Atmosphere Administration, and, in Australia, the Bureau of Meteorology) continue to invest in CGCM development in the expectation that seasonal climate prediction skill will continue to improve.

Despite these strengths, CGCMs have several well-known limitations for streamflow forecasting. CGCMs produce precipitation forecasts that are usually biased at the scale of catchments [e.g. *Crochemore et al.*, 2016], they produce ensembles that often do not represent forecast uncertainty reliably [e.g. *Peng et al.*, 2014], and they sometimes do not simulate the teleconnections between oceans and precipitation as well as statistical methods [*Schepen et al.*, 2012]. These deficiencies can be addressed through post-processing. Some streamflow forecasting methods apply quantile mapping to bias-correct climate forcings and choose to treat deficiencies in reliability by post-processing streamflow forecast outputs [e.g. *Yuan et al.*, 2015b]. Quantile mapping cannot guarantee reliable ensembles [*Wood and Schaake*, 2008], so in these systems the hydrological post-processing usually describes uncertainties of both rainfall and hydrological predictions. We do not follow this approach as it can conflate uncertainties arising from climate and hydrological processes, making it difficult to understand where forecasts can be improved. Instead, we quantify uncertainty in precipitation forecasts separately from uncertainty arising from hydrological modelling.

To generate precipitation forecasts we use the method for calibration, bridging and merging (CBaM) of CGCM forecasts, developed previously [*Schepen and Wang*, 2014; *Schepen et al.*, 2014]. CBaM performs three functions that are basic requirements for any method to post-process ensemble CGCM forecasts: i) it reduces overall and conditional biases, ii) it corrects deficiencies in ensemble reliability and iii) it maximizes the useful signal in forecasts by removing noise.

Land surface processes are very often a crucial source of streamflow forecast predictability. This paper focusses on small (<500 km²) snow-free catchments, and we summarize land surface processes of interest – chiefly soil moisture and groundwater states and fluxes - with the term *catchment wetness*. Catchment wetness tends to have much greater persistence than atmospheric processes, and in many catchments is a more important source of streamflow predictability than climate, particularly at shorter lead-times [e.g. *Bennett et al.*, 2014; *Mahanama et al.*, 2012]. Streamflow responses to initial catchment wetness, together with forecast rainfall, can be effectively simulated with a conceptual hydrological model [e.g. *Day*, 1985; *Robertson et al.*, 2013; *Svensson et al.*, 2015], and we pursue this approach in this study.

The quantification of uncertainty in hydrological modelling and forecasting is an active area of research, with a number of new methods published in recent years. These range from attempts to understand and quantify each source of uncertainty [such as measurement of observations, forcings, model structure, model parameters and optimization strategies; e.g. *Kavetski et al.*, 2006; *Kuczera et al.*, 2006; *Zappa et al.*, 2011]; to methods that focus only on the behavior of residuals [*Evin et al.*, 2014; *Schaefli et al.*, 2007; *Smith et al.*, 2015]. In practice, it is difficult to identify and quantify all sources of hydrologic uncertainty in such a way that they sum to explain the behavior of residuals [*Renard et al.*, 2010]. In addition, methods that focus on residuals have the virtue of (relative) simplicity, and require fewer data. While we argue that it is helpful to distinguish between climate and streamflow forecast uncertainty, we do not believe it is necessary to identify further the sources of hydrological uncertainty. We concentrate on characterizing the behavior of residuals to summarize all sources of hydrological uncertainty, as this affords a more manageable forecast system.

Hydrological error models generally acknowledge that model residuals are heteroscedastic, not normally distributed, and autocorrelated [*Bates and Campbell*, 2001; *Evin et al.*, 2014; *Schaefli et al.*, 2007; *Smith et al.*, 2015]. This often requires complex likelihood functions with which to estimate parameters, leading to undesirable interaction between hydrological and error model parameters [*Evin et al.*, 2013; *Li et al.*, 2015]. One solution is to break down parameter estimation into stages, fixing parameters at each stage before continuing onto the next [*Evin et al.*, 2014; *Li et al.*, 2016]. This approach allows (relatively) simple estimation procedures at each stage, and guards against undesirable parameter interaction. We pursue a staged error model here, incorporating i) the need for forecasts to transit to stochastic scenarios when they are not skillful and ii) our previous finding that varying the error model by season can lead to substantial improvements in characterizing error for monthly streamflow simulations [*Li et al.*, 2013].

FoGSS is a new forecasting system that combines a series of existing and new statistical techniques to link CGCM forecasts to a conceptual hydrological model and to represent uncertainties in rainfall and hydrological forecasts with an ensemble. In this study, we aim to

1) Describe the theoretical basis for the FoGSS model, and
2) Demonstrate its ability to generate accurate, reliable streamflow forecasts to long lead-times in a case study.

The paper is structured as follows. The FoGSS model is described in Section 2 and validation methods are explained in Section 3. Case study catchments are detailed in Section 4 and the results of the case studies are described in Section 5. Implications of the results are discussed in Section 6, and the study is summarized and concluded in Section 7.

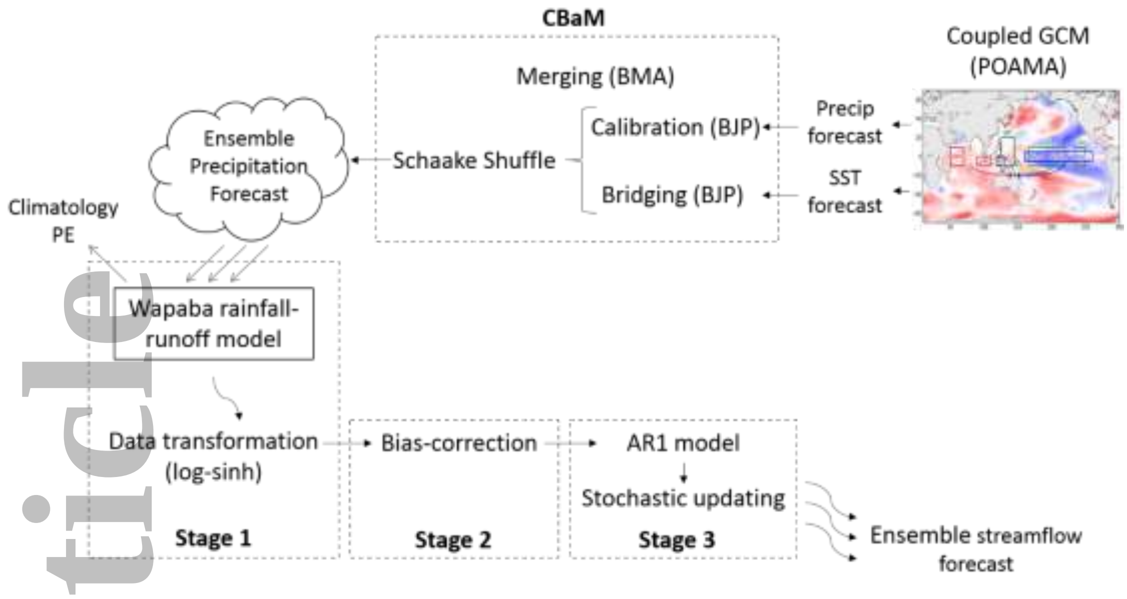


Figure 1 Schematic of the FoGSS model. See text for details.

2. METHODS

To assist the reader through the description of the components of FoGSS, a conceptual representation of the model is shown in Figure 1.

2.1. Ensemble rainfall forecasts: calibration, bridging and merging (CBaM) of CGCM forecasts

Ensemble rainfall forecasts are generated by post-processing forecasts from a CGCM using CBaM. The CGCM we use is the Predictive Ocean and Atmosphere Model for Australia [POAMA version M2.4, *Hudson et al.*, 2013; *Marshall et al.*, 2014]. POAMA produces forecasts to lead-times of 8 months at a monthly time step, using perturbed initial conditions and differing model physics to generate a 33-member ensemble at a grid resolution of 2.5°. (Note that here and throughout the paper, lead-times are zero based: a forecast for the next month has a lead-time of zero, and a time series forecast for the coming year extends from a lead-time of zero to a lead-time of 11 months.) POAMA forecasts are issued at the start of each month (12 forecasts per year). CBaM has been developed and tested in a series of previous papers [*Hawthorne et al.*, 2013; *Peng et al.*, 2014; *Schepen and Wang*, 2014; *Schepen and Wang*, 2015; *Schepen et al.*, 2014; *Wang et al.*, 2012a], and only an overview is presented here. For more information we refer readers to *Schepen and Wang* [2014], which gives a detailed description of the version of CBaM used here, including a full list of equations.

Calibration

To calibrate POAMA rainfall forecasts, we apply the Bayesian joint probability modelling approach [BJP, *Wang and Robertson*, 2011; *Wang et al.*, 2009]. The BJP establishes a joint probability distribution between the ensemble mean of POAMA rainfall forecasts and observations through these steps:

1) Observed and forecast rainfall data are transformed with the log-sinh transformation [*Wang et al.*, 2012b; and see Section 2.2, below] to normalize data and homogenize variance

2) Transformed observed and forecast rainfall data are assumed to follow a bivariate normal distribution

3) The joint posterior distribution of the parameters of the log-sinh transformation and the bivariate normal distribution are inferred using Markov chain Monte Carlo techniques; zero values are treated as censored data in the inference [see *Wang and Robertson*, 2011, and Section 2.2, below]

4) Calibrated rainfall forecasts are generated by drawing 1000 samples from the posterior predictive distribution, which is conditioned on the mean of the POAMA ensemble forecast.

Bridging

To increase forecast skill, we also produce ‘bridging’ forecasts. Bridging takes advantage of skill that is sometimes present in sea-surface temperature (SST) forecasts, due in part to SST’s greater persistence compared to rainfall. In some cases it is possible for statistical techniques to represent teleconnections between SST and precipitation better than the dynamical atmospheric components of CGCMs. Bridging forecasts are produced from the mean of POAMA ensemble forecasts of 6 indices derived from SST (Table 1). Bridging follows the same process as calibrating rainfall forecasts, except that POAMA forecasts of the SST indices are used in place of POAMA forecasts of rainfall. This requires one small adjustment at step one, above: each SST index is transformed with the Yeo-Johnson transformation [*Yeo and Johnson*, 2000], as it can be applied to variables that take both positive and negative values. To extend rainfall forecasts from lead-times of 8 to 11 months, bridging forecasts are produced at lead-times of 9, 10 and 11 months from POAMA forecasts of SST indices at the 8-month lead-time.

Table 1 Sea surface temperature indices used to produce bridging forecasts

Index	*Definition*
NINO3	Average over 150°W–90°W, 5°S–5°N
NINO3.4	Average over 120°W–170°W, 5°S–5°N
NINO4	Average over 160°E–150°W, 5°S–5°N
EMI [El Nino Modoki; *Ashok et al.*, 2007]	C – 0.5(E + W), where C = average over 165°E–140°W, 10°S–10°N E = average over 110°W–70°W, 15°S–5°N W= average over 125°E–145°E, 10°S–20°N
DMI [Indian Ocean Dipole Mode Index; *Saji et al.*, 1999]	WPI – EPI, where EPI = average over 90°E–110°E, 10°S–0°S WPI = average over 50°E–70°E, 10°S–10°N
II [Indonesian Index; *Verdon and Franks*, 2005]	Average over 120°E–130°E, 10°S–0°S

Merging

Calibration and bridging yield seven different ensemble rainfall forecasts: one calibrated rainfall forecast, and six bridging forecasts. The seven forecasts are merged with Bayesian model averaging [BMA; *Wang et al.*, 2012a]. The BMA algorithm weights forecasts by their ability to forecast rainfall under cross-validation. If no single forecast is clearly superior, the BMA is designed to give approximately equal weights to all forecasts.

Schaake shuffle

Thus far we have treated each monthly rainfall forecast as independent from months that precede or succeed it. If temporal patterns in rainfall time series are not realistic, it is difficult to guarantee that total rainfall summed for all lead-times will be realistic. To solve this problem, the Schaake shuffle [*Clark et al.*, 2004] is used to link ensemble members from different lead-times to produce realistic time series forecasts. The Schaake shuffle applies observed temporal rank correlations to the forecast data. In its original conception, the Schaake shuffle matched the number of observations sampled to the size of the ensemble. CBaM generates a 1000-member ensemble, and the observational record is far too short to sample 1000 sequences. We address this problem by separating the ensemble into batches, where each batch is equal in size to the number of observations available. The Schaake shuffle is then applied to each batch.

The result is a 1000-member ensemble rainfall forecast for each month to a lead-time of 11 months. Each ensemble member follows a realistic temporal pattern, and thus each ensemble member can be used to generate a time series streamflow forecast. *Schepen and Wang* [2014] showed that CBaM instills four crucial properties in ensemble rainfall forecasts:

1. Forecasts are unbiased at the catchment scale (i.e., CBaM works very effectively to downscale forecasts to catchments)
2. Forecast ensembles are highly reliable in uncertainty spread
3. Forecasts extract skill available from unprocessed POAMA rainfall and SST forecasts
4. When skill is absent, forecasts converge to climatology.

This last property is a particularly important if the rainfall forecasts are used to produce streamflow forecasts. Because streamflows are often heavily influenced by catchment memory, forecasts of streamflows can be skillful to much longer lead-times than rainfall forecasts [*Shukla and Lettenmaier*, 2011; *Wood and Lettenmaier*, 2008]. But this is true only if the rainfall forecasts do not introduce errors (i.e. negative skill) at longer lead-times. CBaM ensures rainfall forecasts are at least similarly skillful to climatology at all lead-times.

2.2. Hydrological model and error modelling

Overview

A conceptual hydrological model is used to predict the response of runoff to rainfall, and a staged error model is used to refine the statistical properties of the residual in 3 stages:

1) Hydrological modelling and data transformation to normalize data and ensure data are homoscedastic
2) Conditional bias-correction to remove conditional and unconditional bias
3) Autoregressive updating to reduce, as much as possible, the magnitude of errors and to propagate uncertainty through the forecast

A conceptual representation of the stages is shown in Figure 1, and the parameters for each stage are listed in Table 2.

Parameters are estimated using maximum likelihood estimation. We use the shuffled complex evolution algorithm [SCE; *Duan et al.*, 1994] to optimize parameters. Stage 1 parameters are estimated first and then fixed for later stages; Stage 2 parameters are then estimated (and fixed); and then Stage 3 parameters are estimated.

Table 2 Parameters estimated at each stage of the FoGSS hydrological and error model. Parameters followed by (i) vary by month.

Stage	*Model*	*Parameters*
1	log-sinh transformation	a, b, m^*, s^*
	Wapaba	$\alpha_1, \alpha_2, \beta, k, S_{\max}$
	Residual distribution	σ_1^*
2	Bias-correction	$d(i), \mu(i), \sigma_2(i)^*$
3	Restricted autoregressive model	$\rho(i), \sigma_3(i)$

*not used to generate final FoGSS forecast

Stage 1: Hydrological model and data transformation

The water partitioning and balance [Wapaba; *Wang et al.*, 2011] model is used to simulate the response of streamflow to rainfall and potential evaporation. Wapaba is a conceptual hydrological model based on the Budyko curve, which casts the water balance as a competition between available water and available energy. *Wang et al.* [2011] compared Wapaba with several other hydrological models for 331 Australian catchments, and showed that Wapaba performs at least as well as the other models, and often better. Wapaba runs at a monthly time step, and has 5 parameters (Table 2). The function of Wapaba parameters is as follows: $S_{\max}$ controls the size of surface soil moisture stores; α_1 and α_2 control the efficiency with which rainfall and evaporation, respectively, act on the water balance; β controls recharge to groundwater; and k controls discharge from the ground water store. The ground water store is assumed to be of infinite size. To estimate parameters, Wapaba is forced with observed rainfall and potential evaporation (Section 4); i.e., parameters are not estimated from forecasts. An example parameter set for each catchment is given in Appendix A.

<u>Estimating parameters of the log-sinh transformation</u>

Streamflow simulations are transformed with the log-sinh transformation [*Wang et al.*, 2012b] in order to make residuals normally distributed and homoscedastic. The log-sinh transformation has been shown to outperform comparable transformations for hydrological data [*Del Giudice et al.*, 2013; *Wang et al.*, 2012b], and it is also effective at transforming extreme values in hydrological data [*Bennett et al.*, 2014].

The log-sinh transformation is given by

$$z = TF(q) = \frac{1}{b}\log\left(\sinh\left(a + bq\right)\right) \quad (1)$$

where a and b are parameters and q is streamflow. The inverse transformation (back transformation) is given by

$$q = TF^{-1}(z) = \max\left(\frac{1}{b}\left(\arg\sinh\left(e^{bz}\right) - a\right), 0\right), \quad (2)$$

where z is any streamflow that has previously been transformed with Equation (1). For clarity, we will refer to the domain in which q exists as the *original domain* to differentiate it from the *transform domain* of z.

To estimate the values of a and b, Equation (1) is applied to observed streamflows, q_o, to gain transformed observed streamflows, z_o. We assume that z_o follows a normal distribution:

$$z_o \sim N\left(m, s^2\right). \quad (3)$$

The parameters that require estimation are denoted by

$$\pi = \{a, b, s, m\}. \quad (4)$$

The transformation is fitted to observations by maximizing the likelihood

$$p\left(\pi \mid q_o(t)\right) = \prod_{t=1}^{T} J_{z\to q} N\left(z_o(t) \middle| m, s^2\right) \quad (5)$$

where t is each time in the estimation period T, and the Jacobian, $J_{z\to q}$, is given by

$$J_{z\to q} = \frac{1}{\tanh\left(a + bq_o\left(t\right)\right)}. \quad (6)$$

Maximum likelihood estimation assumes independent (e.g., not autocorrelated) and identically distributed data. As already noted, streamflow data is usually autocorrelated, which we ignore in Equation (5). Equation (5) is thus a 'pseudolikelihood' rather than a full maximum likelihood estimation [*Cox and Reid*, 2004]. Parameters related to autocorrelation do not affect the shape of the marginal distribution, and thus can be considered 'nuisance' parameters that may be ignored.

In catchments where zero flows occur, it is highly likely that our assumption of normality (Equation 3) will be violated. To avoid this problem, we assume that the data can be characterized by a normal distribution that encompasses negative values, and we treat zero values as censored data when fitting this distribution [after *Wang and Robertson*, 2011]. Each zero observation is treated as censored data by assuming that i) each zero observation has some unknown value below or equal to zero, and ii) a substitute value can be generated from the normal distribution we are fitting. Mathematically, when $q_o\left(t\right)=0$, we replace the term $J_{z\to q}N\left(z_o(t)\,|\,m,s^2\right)$ in Equation (5) with

$$\Phi\left(\frac{z_c-m}{s}\right), \tag{7}$$

where Φ is the cumulative distribution of a standard normal distribution, and z_c is the log-sinh transformed (Equation 1) value of zero (i.e., $z_c=a^{-1}\log\left(\sinh\left(b\right)\right)$.

Once parameters in π are estimated, we fix a and b for subsequent stages, and discard m and s.

Estimating Wapaba Parameters

To estimate Wapaba parameters we first transform both simulated streamflow, q_s, and observed streamflow, q_o, to z_1 (subscripts of z to denote the stage) and z_o, respectively, using Equation (1). We assume that the residual, ε_1 (where the subscript again denotes stage), in the transform domain is normally distributed:

$$\begin{aligned} z_o\left(t\right)&=z_1\left(t\right)+\varepsilon_1 \\ \varepsilon_1&\sim N\left(0,\sigma_1^2\right) \end{aligned}, \tag{8}$$

with σ_1 the standard deviation of the residual at all time steps, t.

The Wapaba parameters (Table 2) and the fixed log-sinh parameters are denoted as

$$\theta_1=\left\{\alpha_1,\ \alpha_2,\ \beta,\ K,\ S_{\max},\ a,\ b,\sigma_1\right\}. \tag{9}$$

To estimate the Wapaba parameters, the pseudolikelihood

$$p\left(q_o\left(t\right)|\,\theta_1,q_s\left(t\right)\right)=\prod_t^T J_{z\to y}N\left(z_o(t)\middle|z_1(t),\sigma_1\right) \tag{10}$$

is maximized, where the Jacobian is as defined in Equation (6). As with the log-sinh parameter estimation above, instances where $q_o\left(t\right)=0$ are treated as special cases in the likelihood, and the term $J_{z\to y}N\left(z_o(t)\middle|z_1(t),\sigma_1\right)$ in Equation (10) is replaced by $\Phi\left(\dfrac{z_c-z_1(t)}{\sigma_1}\right)$

.

Stage 2: Conditional Bias-correction

In earlier work [*Li et al.*, 2013], we showed that for monthly streamflow simulations it was beneficial to account for month-to-month variation in bias [see also *Yuan*, 2016, who used a similar approach]. We extend this approach here by applying a new bias-correction method that varies between months. We term this method *conditional bias-correction* as it corrects biases occurring in different hydrological conditions.

Bias is corrected in the transformed simulated streamflows at each t by

$$z_2(t) = d(i)\, z_1(t) + \mu(i), \tag{11}$$

where $d(i)$ and $\mu(i)$ are parameters that vary by month

$$i = 1, 2, ..., 12 = month(t). \tag{12}$$

As with Stage 1, we assume that residuals are normally distributed. Because parameters in Equation (11) vary by month it follows that the residual distribution also varies by month:

$$\begin{aligned} z_o(t) &= z_2(t) + \varepsilon_2(i) \\ \varepsilon_2(i) &\sim N\left(0, \sigma_2^2(i)\right) \end{aligned} \tag{13}$$

The conditional bias-correction in Equation (11) takes the form of a simple linear regression, but because it is applied to transformed data it can correct strongly non-linear and conditional biases in q. When d is at or near zero, the conditional bias-correction effectively ignores simulated streamflows generated at Stage 1, and $z_2 \approx \mu$. That is, the conditional bias-correction performs the important function of forcing simulations to a constant (akin to climatology) in months where transformed they perform poorly.

The Stage 2 parameters to be estimated are denoted as

$$\theta_2(i) = \{d(i), \mu(i), \sigma_2(i)\} \tag{14}$$

To estimate Stage 2 parameters we maximize the pseudolikelihood

$$p\left(q_o(t) \middle| \theta_2(i), q_1(t)\right) = \prod_{t \in T_i} J_{q \to z} N\left(z_o(t) \middle| z_2(t), \sigma_2(i)\right). \tag{15}$$

Where q_1 is the simulation produced with Stage 1. As already noted, the estimation is performed separately for each month: T_i is the set of all observations for a given month i within the estimation period.

As when estimating log-sinh parameters in Stage 1, if $q_o(t)=0$ then the term $J_{q\to z}N\left(z_o(t)\middle|z_2(t),\sigma_2(i)\right)$ in Equation (15) is substituted with $\Phi\left(\frac{z_c-z_2(t)}{\sigma_2(i)}\right)$.

Stage 3: Autoregressive updating

Streamflow tends to be highly autocorrelated, and it follows that residuals of streamflow simulations tend to be autocorrelated [*Bates and Campbell*, 2001; *Schoups and Vrugt*, 2010; *Smith et al.*, 2015]. Once transformed flows have been bias-corrected, a first-order autoregressive (AR) model is applied to account for autocorrelation in the residuals. The AR model takes the form

$$z_3(t)=z_2(t)+\rho(i)\left(z_o(t-1)-\max\left(z_2(t-1),z_c\right)\right). \tag{16}$$

As with Stage 2, ρ varies with month following earlier work [*Li et al.*, 2013] that showed that it was beneficial to vary autoregression coefficients by month for monthly rainfall-runoff modelling.

We make one exception to the exact form of Equation (16). To prevent updating by too large an amount, the restricted AR model proposed by *Li et al.* [2015] is used. To apply the restriction, we first back-transform z_2 to q_2 and z_3 to q_3 with Equation (2). After applying Equation (16), we check if

$$\left|q_3(t)-q_2(t)\right|>\left|q_o(t-1)-q_2(t-1)\right|. \tag{17}$$

If this is true, q_3 is revised to

$$q_3(t)=q_2(t)+q_o(t-1)-q_2(t-1) \tag{18}$$

and $q_3(t)$ is transformed to $z_3(t)$ using Equation (1). That is, the update applied in Equation (16) is never allowed to exceed the residual in the original domain at the previous time step. The restriction is necessary because when flows are rising, the transformation can amplify an update to become unrealistically large in the original domain [*Li et al.*, 2015]. When flows are falling, the tendency of the transformation to reduce the update with the magnitude of flows is highly beneficial.

As with previous stages, we assume that Stage 3 residuals are normally distributed:

$$\begin{aligned} z_o(t)&=z_3(t)+\varepsilon_3(i)\\ \varepsilon_3(i)&\sim N\left(0,\sigma_3^2(i)\right) \end{aligned}. \tag{19}$$

The parameters to be estimated at Stage 3 are denoted as:

$$\theta_3(i)=\{\rho(i),\sigma_3(i)\} \ . \tag{20}$$

As this is the final stage, the standard deviation of the residual (σ_3) is not discarded. To estimate Stage 3 parameters, we maximize the likelihood

$$p\left(q_o(t)\middle|\theta_3(i),q_o(t-1),q_2,q_2(t-1)\right)=\prod_{t\in T_i} J_{q\to z} N\left(z_o(t)\middle|z_3(t),\sigma_3(i)\right). \tag{21}$$

Unlike previous stages, Equation (21) is not a pseudolikelihood but rather a full maximum likelihood estimation that treats residuals as independent and identically distributed.

Finally, and as for previous stages, if $q_o(t)=0$ then the term $J_{q\to z} N\left(z_o(t)\middle|z_3(t),\sigma_3(i)\right)$ in Equation (21) is substituted with $\Phi\left(\dfrac{z_c-z_3(t)}{\sigma_3(i)}\right)$.

Only the standard deviation at Stage 3, σ_3, is used to generate the final FoGSS forecast. The standard deviations at stages 1 and 2, σ_1 and σ_2, are usually discarded. In this study we use σ_1 and σ_2 to demonstrate the function of each stage (see Section 3).

2.3. Generating a forecast

We now have a collection of parameters, $\theta=\{\theta_1,\theta_2,\theta_3\}$, from which to generate a forecast. A streamflow forecast is generated for each rainfall forecast ensemble member. Before a forecast is issued, Wapaba model states are warmed up by forcing the model with observations for the 10-year period immediately preceding the forecast issue time. An initial streamflow forecast is generated by forcing Wapaba with one forecast rainfall ensemble member and climatology potential evaporation (PE). Climatology PE is a satisfactory forcing, as streamflow forecasts are not sensitive to PE in comparison to rainfall. The initial forecast is transformed and bias-corrected with the first two stages of the error model (equations 1 and 11, respectively), and we denote the transformed and bias-corrected forecast as z_{F2}, and its back-transform (Equation 2) as q_{F2}.

Stage 3 performs two functions in generating a forecast: i) correcting errors at shorter lead-times with autoregressive updating, ii) quantifying and propagating uncertainty through all lead-times. To perform these functions, we use a technique called *stochastic updating* to apply the restricted AR model in Stage 3 [*Wang et al.*, 2016]. In essence, stochastic updating substitutes observed flows in equations 16-18 with the most recent updated forecast. This substitution is similar in concept to that used in the statistical post-processor proposed by *Seo et al.* [2006], although the structure of our error model is substantially different. For a forecast issued at time t and at lead-time $\tau = 0,1,\ldots,11$,

$$z_{F3}\left(t+\tau\right) = z_{F2}\left(t+\tau\right) + \rho\left(i_F\right)\left(z_o\left(t+\tau-1\right) - \max\left(z_{F2}\left(t+\tau-1\right), z_c\right)\right) \qquad (22)$$

(essentially identical to Equation 19), where $i_F = 1,2,\ldots,12 = month(t+\tau)$. $z_{F3}\left(t+\tau\right)$ is back-transformed (Equation 2) to $q_{F3}\left(t+\tau\right)$, and the restriction is applied, following Equation (17) and Equation (18):

$$z_{F3}\left(t+\tau\right) = \begin{cases} z_{F3}\left(t+\tau\right) & \left(\left|q_{F3}\left(t+\tau\right) - q_{F2}\left(t+\tau\right)\right| \le \left|q_o\left(t+\tau-1\right) - q_{F2}\left(t+\tau-1\right)\right|\right) \\ TF\left(q_{F2}\left(t+\tau\right) + q_o\left(t+\tau-1\right) - q_{F2}\left(t+\tau-1\right)\right) & \left(\left|q_{F3}\left(t+\tau\right) - q_{F2}\left(t+\tau\right)\right| > \left|q_o\left(t+\tau-1\right) - q_{F2}\left(t+\tau-1\right)\right|\right) \end{cases},$$

(23)

where TF is the log-sinh transformation given by Equation (1).

To generate a streamflow forecast ensemble member in the transform domain, z_F, a noise component is added

$$z_F\left(t+\tau\right) = z_{F3}\left(t+\tau\right) + \varepsilon_F\left(i_F\right), \qquad (24)$$

where ε_F is a single random sample drawn from $N\left(0, \sigma_3^2\left(i_F\right)\right)$. Finally, z_F is back-transformed to q_F with Equation (2).

At the first lead-time $\tau = 0$, $q_o\left(t-1\right)$ and $z_o\left(t-1\right)$ are available. At each succeeding lead-time, $\tau = 1,2,\ldots,11$, $z_F\left(t+\tau-1\right)$ substitutes for z_0 in Equation (22), and $q_F\left(t+\tau-1\right)$ substitutes for q_o in Equation (23). Uncertainty is then propagated through each lead-time by equations 22-24.

This process is repeated for each rainfall forecast ensemble member. The large size of the rainfall forecast ensemble (1000 members) means that we only need to generate one streamflow forecast ensemble member for each rainfall forecast ensemble member.

Forecasts are generated at the end of each calendar month – that is, we generate 12 forecasts each year.

3. MODEL AND FORECAST VALIDATION

To demonstrate the function of each stage of the hydrological and error model, we validate forecasts generated after Stage 1, Stage 2 and Stage 3. To generate Stage 1 and Stage 2 forecasts the standard deviations σ_1 and σ_2, respectively, are used in the error models. As already noted, σ_1 and σ_2 would ordinarily be discarded and not used in the full FoGSS model.

3.1. Statistical assumptions of the error model

Before evaluating forecast skill, we test the statistical assumptions we make about residuals in the error model. For these tests we are interested in the theoretical ability of the model to describe residuals, so parameters are estimated using all data available for the period 1982-2009 (i.e., without cross-validation). With these parameters we generate lead-0 'predictions' for the 1982-2009 period by forcing Wapaba with observed rainfall and PE, and calculate prediction residuals in the original domain, denoted ε_q. The three stages of the error model are then applied, and residuals of the median prediction in the transform domain are calculated for each stage, denoted ε_1, ε_2 and ε_3. As described in Section 2.2 above, at Stage 2 and Stage 3 we assume that residuals follow normal distributions with different variances for each month. To ensure that Stage 2 and Stage 3 residuals can be compared across months, ε_2 and ε_3 are standardized by dividing by σ_2 and σ_3, respectively. We use normal probability plots to check for normality of residuals, scatter plots of residuals to test heteroscedasticity, and the autocorrelation function to analyze autocorrelation of residuals.

3.2. Forecast cross-validation

Forecasts are validated over the 28-year period 1982-2009. This period is used both to estimate parameters and to validate forecasts, using cross-validation. Different cross-validation schemes are used to generate rainfall forecasts with CBaM than for the hydrological component of FoGSS. CBaM parameters are inferred with a leave-3-years-out scheme. When generating forecasts for a given year (e.g., 1990), data from the 3-year period centered on that year (1989-1991) are omitted when inferring CBaM parameters. The climatology potential evaporation used to generate streamflow forecasts is calculated by taking the mean of monthly PE for the period 1982-2010, using the same leave-3-years-out cross-validation scheme as for rainfall.

For the hydrological component of FoGSS, data from the 5-year period that begins at the start of the target year are omitted (e.g., to verify forecasts for 1990, data from 1990-1994 are omitted). The more stringent cross-validation scheme applied to the hydrological component of FoGSS is necessary to account for the strong influence of catchment memory on streamflow, and is used for other operational systems in Australia [e.g. *Lerat et al.*, 2015]. In addition, when estimating Wapaba parameters and generating forecasts, the states of the Wapaba model are first warmed up by running the model from 1970.

3.3. Validation scores and reference forecasts

To measure the accuracy of forecasts we use the well-known continuous ranked probability score (CRPS). CRPS measures the error of all ensemble members with respect to observations by integrating the squared distance between forecast and observed cumulative distribution functions [see, e.g., *Gneiting and Katzfuss*, 2014]. Skill is measured against climatology. Forecast skill is given by the continuous ranked probability skill score (CRPSS):

$$CRPSS = \frac{CRPS_{\mathrm{Ref}} - CRPS_F}{CRPS_{\mathrm{Ref}}} \times 100\%, \qquad (25)$$

where $CRPS_F$ and $CRPS_{\mathrm{Ref}}$ are CRPS values for FoGSS and cross-validated climatology forecasts, respectively. To generate climatology reference forecasts, a log-sinh transformed (Equation 1) normal distribution is fitted to the data using the BJP. 1000 samples are drawn from the transformed normal distribution to generate a climatology. Climatology is generated using observed data for the period 1980-2009, applying the same leave-5-years-out cross-validation procedure as described for the hydrological modelling (Section 3.2). In addition, the statistical significance of skill is tested by generating a sample distribution of CRPSS by bootstrapping with 1000 repeats. When 97.5% of the distribution exceeds zero, we conclude that forecast skill is significant; when no more than 2.5% of the distribution exceeds zero we conclude that forecast skill is significantly negative.

We evaluate the relative bias of forecasts at each lead-time, τ, by

$$Bias(\tau) = \frac{\overline{q_F(\tau)} - \overline{q_o}}{\overline{q_o}} \times 100\%, \qquad (26)$$

where $\overline{q_F(\tau)}$ is the mean of all ensemble forecasts at each lead-time. (For brevity, relative bias is simply referred to as bias throughout the paper.) Biased forecasts may not be compatible with, e.g., storage operation planning models run by water agencies.

The reliability of the ensemble is checked with probability integral transform (PIT) uniform probability plots. Given the cumulative distribution function of the forecast, $C(q_F)$, the PIT of an observed value, q_o, is given by:

$$PIT = C\left(q_o\right) \tag{27}$$

For reliable forecasts, the PIT is uniformly distributed between 0 and 1. That is, if PIT values are plotted against a standard uniform variate, reliable forecasts will follow the diagonal. This is the basis of PIT uniform probability plots (shortened to *PIT plots*).

4. CASE STUDIES AND DATA

The performance of FoGSS is assessed for two Australian catchments, the tropical Barron River catchment and the upper reaches of the temperate Yarra River catchment (Figure 2). The two catchments feed the water supply reservoirs Lake Tinaroo and the Upper Yarra Reservoir, respectively. For the Barron River, flows are forecasted at the gauge just above Lake Tinaroo (Barron River at Picnic Crossing, gauge number 110003A, catchment area 228 km²). For the Upper Yarra Reservoir a high quality inflows series accounting for all inflows from the 340 km² catchment is available (calculated from reservoir levels, outflows and gauges). Flow data for both catchments were supplied by the Bureau of Meteorology. The two catchments have strikingly different climates (Figure 2). The Barron experiences monsoonal rainfalls from December to March, with peak flows in February and March. The upper Yarra receives most rainfall in the austral winter to early spring (June–October) with a relatively dry summer, and streamflow is accordingly highest in late winter and early spring.

Rainfall and potential evaporation (PE) observations are taken from the gridded Australian Water Availability Project dataset [AWAP, http://www.bom.gov.au/jsp/awap; *Raupach et al.*, 2008], which interpolates gauged observation to a 0.05 ° grid. AWAP PE is calculated using the Priestley-Taylor method, which relies on solar irradiance [*Priestley and Taylor*, 1972]. Catchment rainfall and PE are obtained by area-weighted averaging of AWAP grid cells that intersect with the catchment.

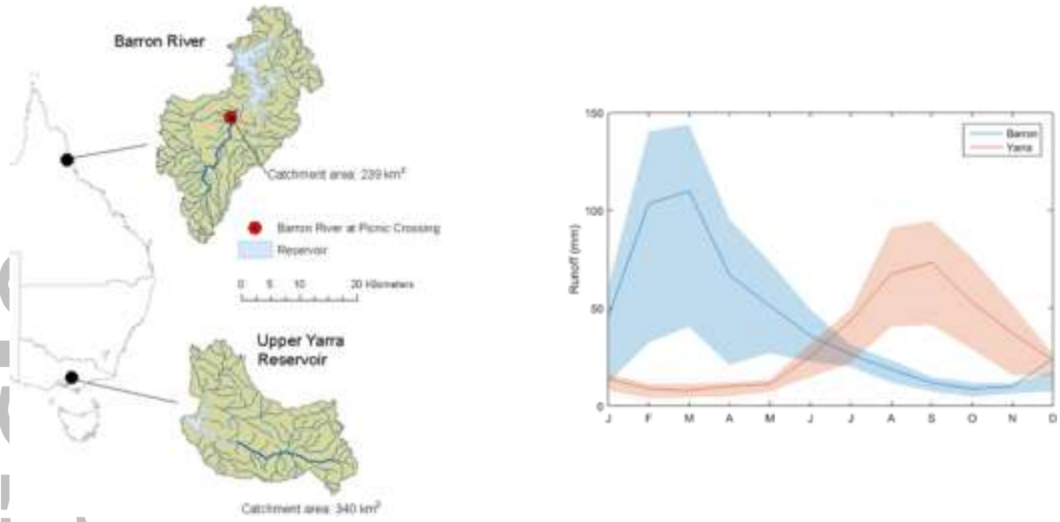


Figure 2 Catchment characteristics. Left panel shows location and catchment areas of case-study sites. Right panel shows runoff characteristics of study catchments; lines show mean runoff, shading gives interquartile range.

5. RESULTS

Examples of retrospective FoGSS forecasts for both catchments are shown in Figure 3. Each ensemble member represents a plausible realization of a 12-month streamflow time series.

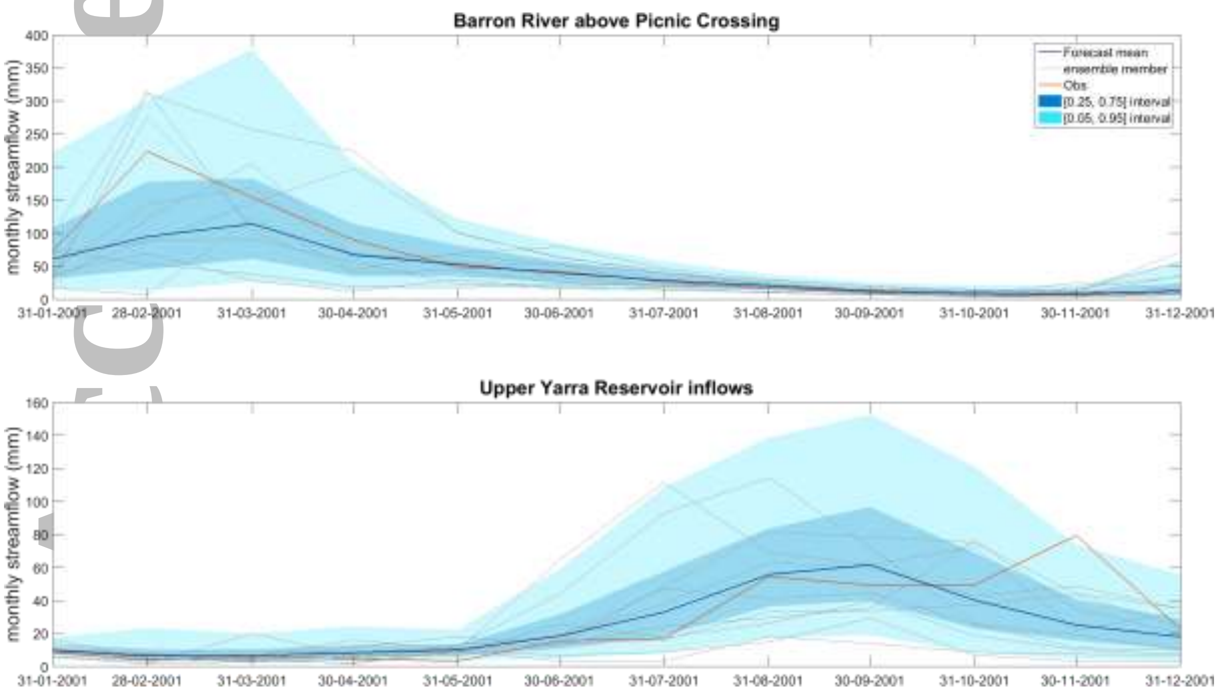


Figure 3 Example retrospective forecasts issued on 31/12/2000 for the Barron River (top) and Yarra River (bottom). Note that only 10 randomly selected ensemble members are displayed of the full 1000-member ensemble.

Table 3 Performance of cross-validated deterministic simulations of the FoGSS hydrological and error model

	Barron River		*Yarra River*	
Metric	*Wapaba only*	*Wapaba + error model*	*Wapaba only*	*Wapaba + error model*
Nash-Sutcliffe efficiency*	0.75	0.87	0.82	0.81
Correlation*	0.87	0.88	0.83	0.83
Bias†	-6.7%	-4.4%	8.7%	1.1%

*for ensemble simulations, calculated from the median of the ensemble
†for ensemble simulations, calculated from the mean of the ensemble

5.1. Statistical characteristics of streamflow residuals

Using traditional deterministic measures of simulation performance, Wapaba simulations are a close match to observations in both catchments, and this performance is improved by the error model (Table 3).

The statistical behavior of residuals is very similar in both catchments, so for brevity we focus on the Barron River. Figure 4 shows the behavior of streamflow residuals (without cross-validation) before the error model is applied (ε_q) and at each stage of the error model (ε_1, ε_2, ε_3). As expected, ε_q is i) not normally distributed, as shown by the departure from the diagonal in the normal probability plot (Figure 4A), and ii) heteroscedastic, as shown by the tendency of errors to get much larger as simulations increase (Figure 4E). Both these problems are addressed by the transformation at Stage 1. Figure 4B-C show that residuals in the transform domain follow a normal distribution for all stages. Figure 4F-H show no clear relationship between residuals and the magnitude of simulations, indicating that they are homoscedastic.

The influence of the bias-correction (Stage 2) on the behavior of residuals shown in Figure 4 is negligible. Nevertheless, Stage 2 plays a crucial role in removing negative forecast skill and correcting biases, as we will show below.

ε_q shows strong autocorrelation at lag-1 (Figure 4I). Autocorrelation of residuals in the transform domain is also strongly prevalent after Stage 1 and Stage 2 (Figure 4J-K), and to much longer lag times. The transformation compresses the differences between large values, leading to higher autocorrelation at longer lag times. Stage 3 strongly reduces autocorrelation in residuals in the transform domain (Figure 4L).

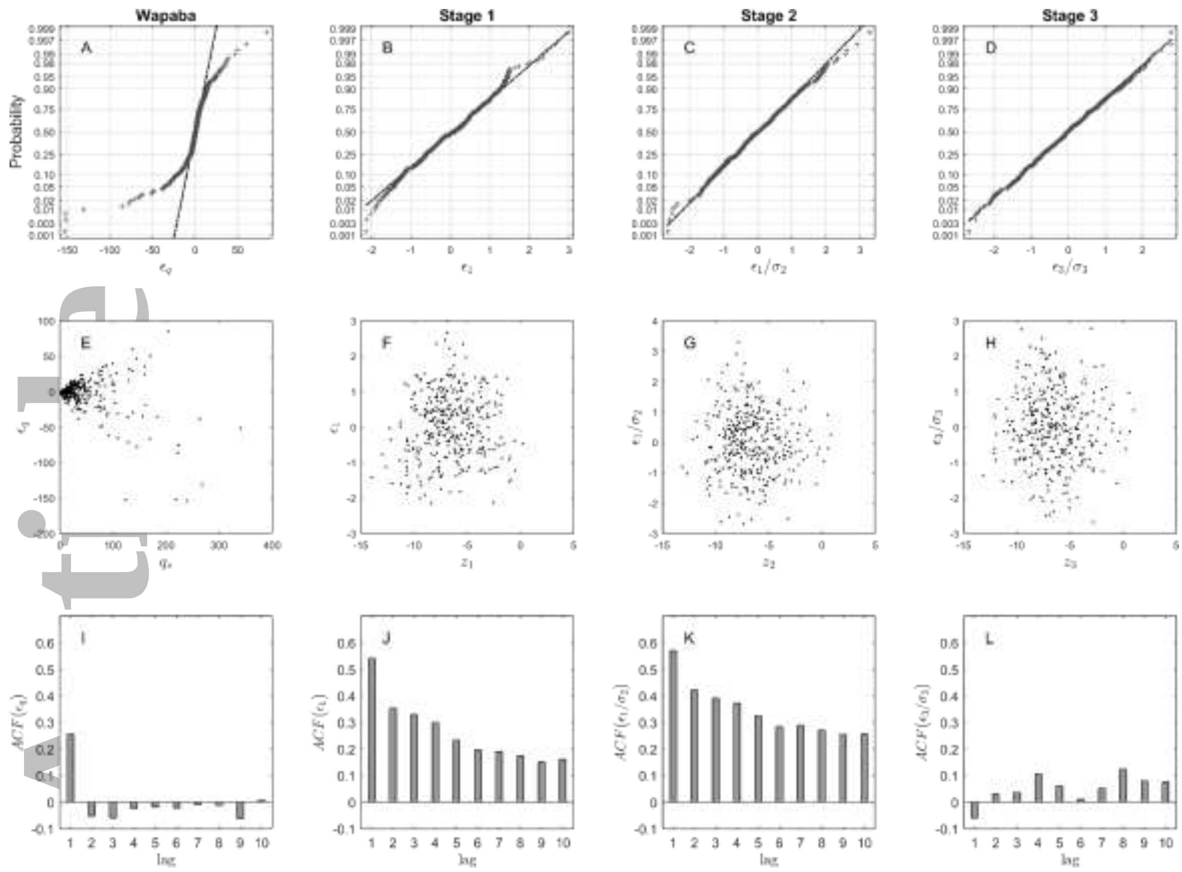


Figure 4 Behaviour of residuals for one-time-step-ahead predictions for the Barron River. Columns show different stages of the FoGSS model: left-most column (A, E, I) shows behaviour of residuals of the Wapaba model in the original domain before error modelling is applied; columns 2-4 show behaviour of residuals in the transform domain after successive stages of the error model are applied (Stage 2 and Stage 3 residuals are standardized to enable comparison of all months). Top row (A-D) gives normal probability plots of residuals (points closer to the dashed black line indicate normality); middle row (E-H) plots residuals against the magnitude of the simulation; bottom row (I-L) shows autocorrelation function (ACF) of residuals.

After Stage 3, the residuals are uncorrelated, normally distributed and heteroscedastic, satisfying the assumptions implicit in Equation (20). Thus we conclude that our model accurately characterizes the behavior of residuals at lead-0 forecasts without cross-validation, when Wapaba is forced by observed rainfall. We now turn our attention to rainfall and streamflow forecasts under cross-validation to long lead-times.

5.2. Rainfall forecast skill

As already noted, the ability of CBaM to produce unbiased, reliable ensemble rainfall forecasts from GCMs is well established, and we do not repeat the in-depth analysis of this technique available elsewhere [*Hawthorne et al.*, 2013; *Peng et al.*, 2014; *Schepen and Wang*, 2013; *Schepen and Wang*, 2014; *Schepen et al.*, 2016]. In particular, we refer readers interested in more detailed analysis of CBaM performance in the Barron catchment to *Schepen et al.* [2018].

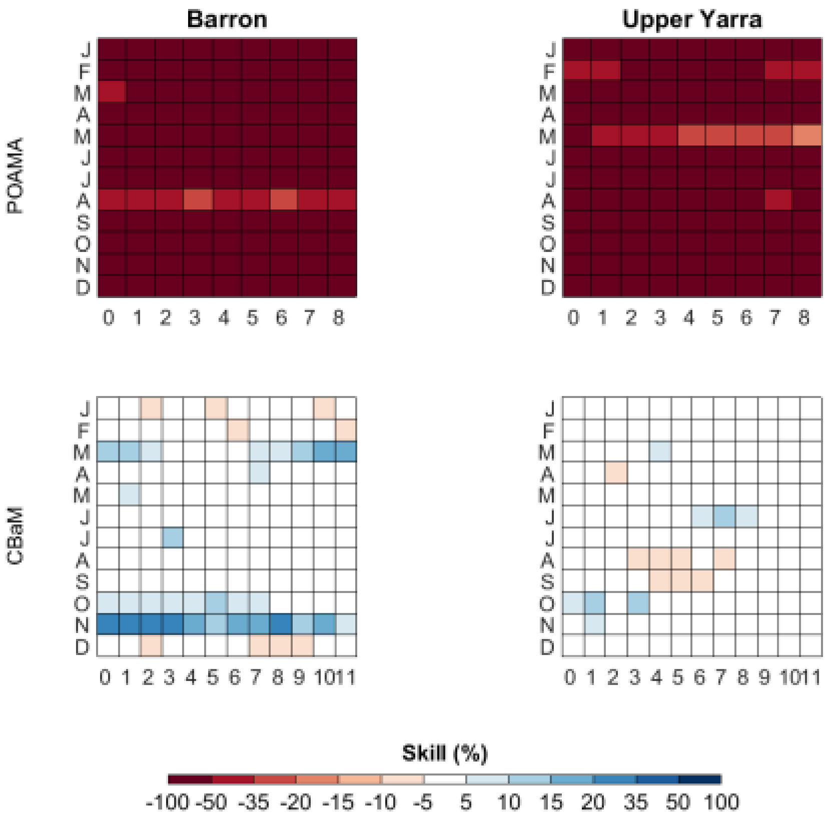


Figure 5 Continuous ranked probability skill score (CRPSS, see text) for precipitation forecast skill from CBaM post-processed POAMA forecasts. In each panel: rows show target month, columns show forecast lead-time.

The improvement in rainfall forecast skill for the Barron and Yarra catchments is shown in Figure 5. Without post-processing, POAMA rainfall forecasts are strongly negatively skillful. In very small catchments with high rainfall, this is unsurprising: POAMA tends to strongly underestimate rainfalls in these catchments (biases of <-50% in both catchments), and the ensemble spread is much too narrow (not shown). In short, raw POAMA forecasts cannot be used to force hydrological models. CBaM improves raw POAMA forecasts dramatically. CBaM rainfall forecasts are largely neutrally skillful (skills around zero) for the majority of months and lead-times in both catchments. That is, CBaM is essentially returning a climatology forecast in most cases. Some skill is present in rainfall forecasts for the Barron catchment in October and November, corresponding to the start of the monsoon-driven wet season, and also in March, which marks the end of the wet season. Of note is the skill in October and November that is available to long lead-times. For the Barron catchment, *Schepen et al.* [2018] showed that skill in November rainfall forecasts is elicited by calibration, while skill in October at longer lead-times is largely contributed by bridging (i.e., from SST forecasts), showing the value of all the components in the CBaM approach.

5.3. Streamflow forecast skill, bias and reliability

Streamflow forecast skill for individual months at all lead-times is given in Figure 6. Forecast skill is high at lead-0, irrespective of the stage of the error model. At lead-0, catchment wetness has a strong influence on forecast skill, and therefore streamflow is reasonably predictable. At longer lead-times, however, forecasts made with Stage 1 can become strongly negative with respect to climatology (e.g. August and September forecasts for the Barron River; November and December forecasts for the Yarra River). The Stage 2 conditional bias-correction plays a crucial role in ameliorating negative forecast skill, all but completely removing negative skills in forecasts of the Yarra River, and substantially removing negative skills for the Barron River. The result is that forecasts effectively default to climatology when skill is not available. Stage 3 improves forecast skill slightly at shorter lead-times (lead-0 and lead-1), with statistically significant lead-0 forecast skill in a majority of months for both catchments. In some months forecasts can be significantly skillful to lead-times >6 months (e.g. November in the Barron catchment, March and February in the Yarra catchment).

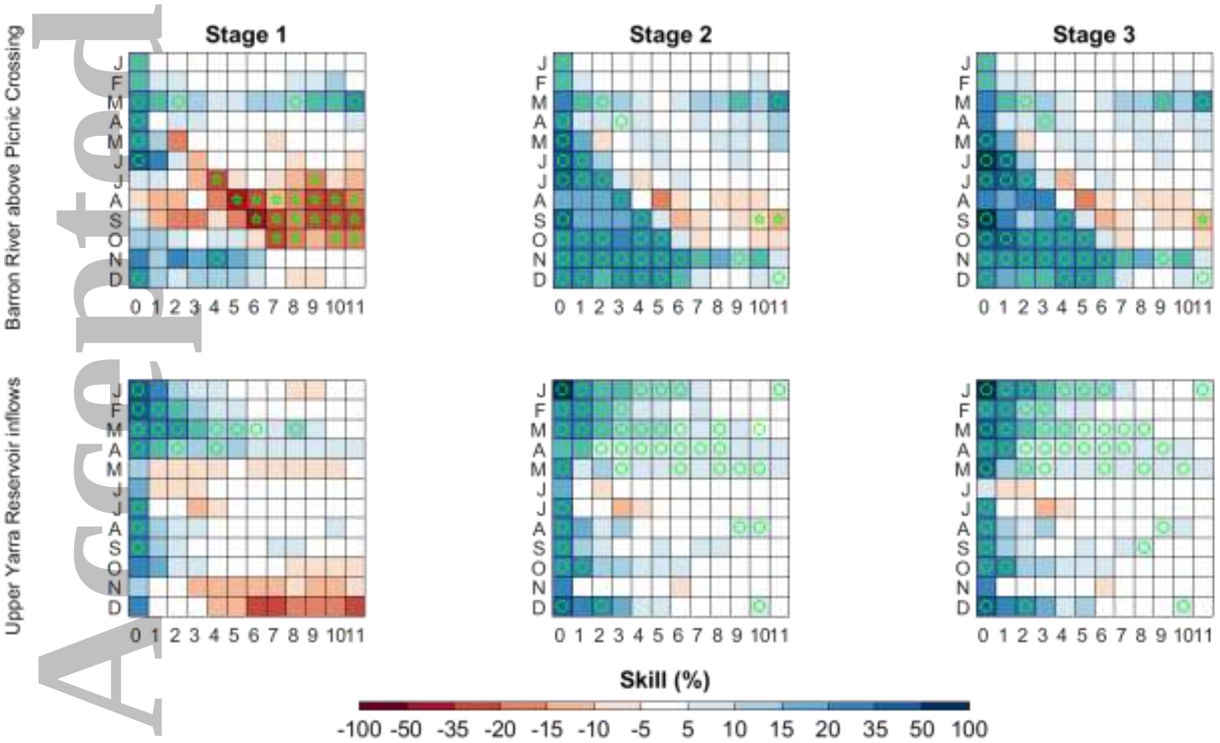


Figure 6 Continuous ranked probability skill score (CRPSS, see text) for each stage of the FoGSS error model. In each panel: rows show target month, columns show forecast lead-time. Top three panels show Barron River, bottom three panels show Yarra River. Green circles show where forecasts are significantly skillful at the 2.5% level; green stars show where forecasts are significantly negatively skillful at the 2.5% level.

Some instances of moderately negative skill persist after Stage 2 in the Barron River (forecasts for August at lead-5 and for September at lead-6 and lead-11). There are two main causes of the negative skill: i) bias in the Wapaba model at lower flows and ii) isolated overestimations of hydrograph recessions in individual years. The first of these manifests in Figure 6 as horizontal lines of negative skill (e.g., for Stage 1 August forecasts); these biases are largely corrected by Stage 2. There are also minor contributions to bias from rainfall forecast bias and the error model (discussed in the next paragraph), which are not corrected by Stage 2. The second cause of negative skill manifests as diagonal lines of negative skill (e.g. the line extending from May, lead-2 to October, lead-7). A major reason for this is the forecast issued at the end of February 1991 (not shown), which significantly overestimated the recession of the annual hydrograph, leading to negative skills from May to October. Because flows are lower in these months, this single event has a substantial impact on relative skill scores. Overall, however, instances of negative skill are very few after Stage 3, of a magnitude we consider acceptable (>-15%), and negative skill is almost never statistically significant, with only one exception (lead-11 September forecasts in the Barron catchment).

Bias is virtually eliminated in the Yarra catchment, while there is a small bias in forecasts for the Barron catchments at longer lead-times (<12%; Figure 7). The conditional bias-correction is more effective in the Yarra catchment because i) bias in Wapaba simulations is higher for the Yarra catchment and thus requires a stronger correction and ii) bias in Stage 1 forecasts changes little with lead-time for the Yarra catchment, but increases with lead-time for the Barron catchment. When bias is low in the simulations, as with the Barron catchment, the bias-correction applied is slight. Thus positive bias in Barron rainfall forecasts, even though it

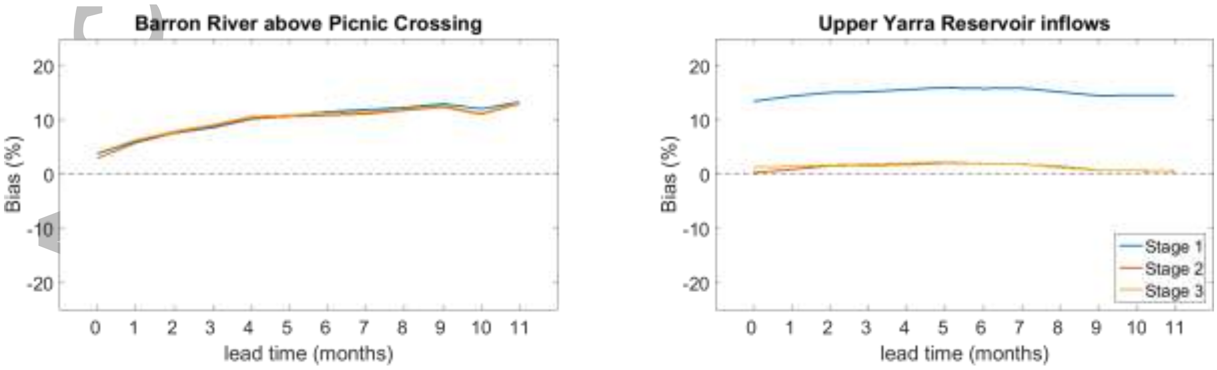


Figure 7 Bias as a function of lead-time. Colours show forecasts generated from the three different stages of the hydrological and error model.

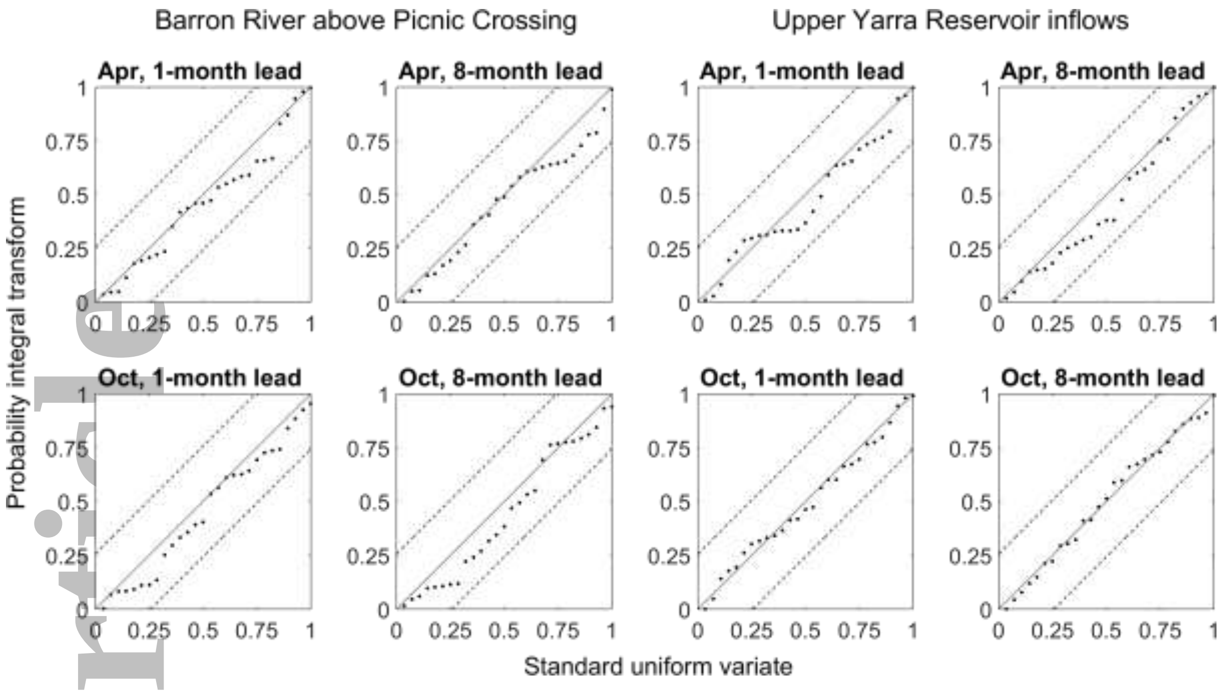


Figure 8 Probability integral transform (PIT) uniform probability plots for Stage 3 forecasts for April and October, at lead-times of 1 and 8 months. Left four panels show Barron River, right four panels show Yarra River. Points give PIT values. Dashed lines give 95% Kolmogorov-Smirnoff confidence intervals. Points close to the diagonal indicate a reliable series of forecasts.

is small (~5%), manifests in the streamflow forecasts. In addition, there is a general tendency of the stochastic updating procedure to induce positive biases at longer lead times. This effect is likely to be small for FoGSS because forecast uncertainty is only propagated to 12 lead-times - we discuss this issue further in Section 6.

Figure 8 shows PIT plots for selected months and lead-times. Since CBaM rainfall forecasts are reliable [*Schepen and Wang*, 2014] and the error model correctly describes the behavior of residuals (Section 0), we expect FoGSS forecasts at short lead-times to be reliable. Figure 8 shows that this is true for lead-1 forecasts. To ensure that forecasts at longer lead-times are also reliable, uncertainty must be correctly propagated through the forecasts. Figure 8 shows that lead-8 forecasts are also reliable, strongly suggesting that the stochastic updating is propagating uncertainty correctly through the forecast. Indeed, forecasts for all months and lead-times are reliable for both catchments (not shown for brevity).

Further evidence for the efficacy of stochastic updating is presented in Figure 9, which shows PIT plots for forecasts of accumulated volumes. It is difficult to achieve reliable forecasts of accumulated volumes if i) streamflow temporal sequences of individual ensemble members are not realistic, or ii) if uncertainty is not correctly propagated through a forecast. Figure 9

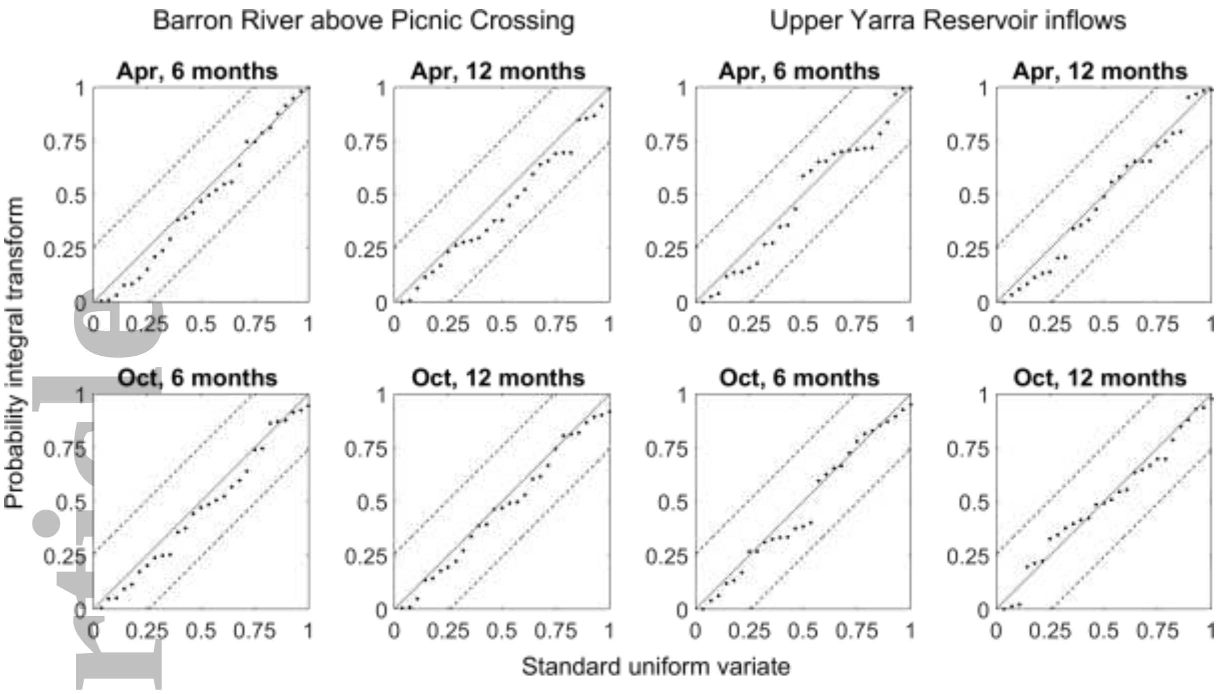


Figure 9 Probability integral transform (PIT) uniform probability plots for Stage 3 forecasts of accumulated flows. Forecasts are accumulated for 3- and 12-month periods for which the first months are April (top row) and October (bottom row). Left four panels show Barron River, right four panels show Yarra River. Points give PIT values. Dashed lines give Kolmogorov-Smirnoff 95% confidence intervals. Points close to the diagonal indicate a reliable series of forecasts.

shows that forecasts of accumulated volumes are highly reliable, giving a strong indication that individual ensemble members have realistic temporal sequences, and that uncertainty is correctly propagated. (Again, this is true for all accumulation periods and months in both catchments, but these are not shown for brevity.)

Achieving skillful forecasts at short lead-times, removing negative forecast skill, and correctly propagating uncertainty through all lead-times are powerful attributes for applications where forecasts of accumulated volumes are important. If these properties are present, skill at short lead-times is able to carry through to longer accumulation periods. Skill for accumulated forecasts is shown in Figure 10. Even though positive skill at individual months typically recedes by lead-3, forecasts of accumulated 6-month streamflow volumes are still strongly skillful. In both catchments, 6-month volume forecasts issued at three different months exhibit statistically significant skill. Forecasts of 12-month volumes are also skillful, although less so, as the influence of skill from early lead-times wanes. Note that the few instances of negative forecast skills in the Barron River do not greatly diminish the skill of forecasts of accumulated 12-month streamflow volumes. This is because instances of negative skill are few, are not strongly negative, and occur in low-flow months, meaning that they are outweighed by positive skill in high-flow months when volumes are accumulated.

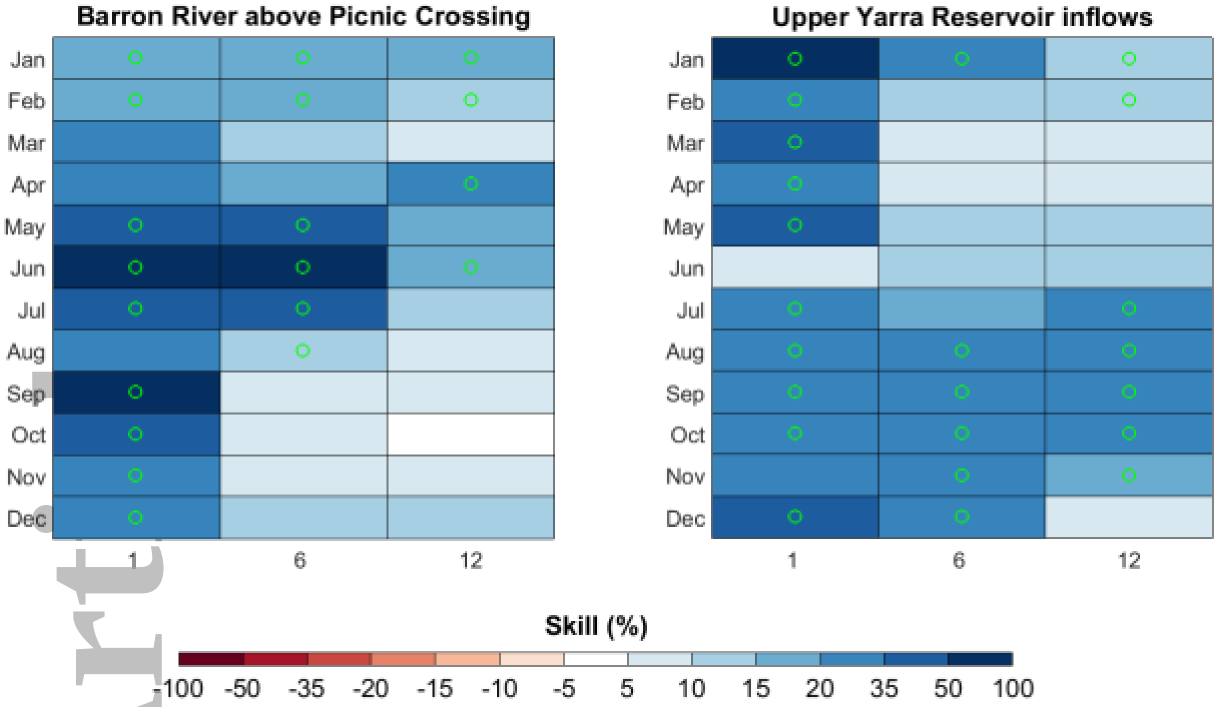


Figure 10 Continuous ranked probability skill score (CRPSS, see text) for accumulated Stage 3 forecasts. Rows show the first month of the accumulation period, columns show forecast accumulation period. Green circles show where forecasts are significantly skillful at the 2.5% level.

6. DISCUSSION

Skillful forecasts of 12-month accumulated volumes are only possible because each element of the FoGSS forecasting system plays its part:

i) CBaM produces highly reliable rainfall forecasts that are as skillful as possible
ii) Wapaba simulates catchment wetness and the runoff response to rainfall
iii) The behavior of streamflow residuals is well characterized by the staged error model
iv) Instances of negative skill are largely eliminated by the conditional bias-correction
v) Hydrological forecast uncertainty is correctly propagated by the restricted autoregressive model and stochastic updating

In essence, skill from shorter lead-times (typically 0-3 months) carries through to forecasts of long accumulation periods because skill at short lead-times is not corrupted by negative forecast skill or by incorrect propagation of forecast uncertainty.

Skill in accumulated volume forecasts can also be present because errors in individual months are cancelled. It is generally more difficult to forecast flow for a particular month than it is to forecast total volumes, because the timing of streamflow peaks and troughs becomes less important for forecasts of accumulated volumes. This explains the interesting result that arises in the Yarra River catchment: skill of forecasts for accumulated 12-month volumes issued at the end of January and June is statistically significant, while 6-month volume forecasts issued at the end of January and June are not significantly skillful (though still positive). South-east Australia suffered one of the worst droughts in recorded history over the period 2001-2009 [*van Dijk et al.*, 2013]. The persistent low streamflows during this period dramatically reduced 12-month streamflow volumes from the upper Yarra River. The FoGSS model was better able to represent these prolonged periods of low flow than climatology forecasts under the leave-5-years out cross-validation scheme, resulting in significantly positive skill for 12-month volume forecasts. Conversely, forecast skill of 6-month volumes reflects how accurately we forecast the onset of wetter conditions in autumn or the onset of drier conditions in spring/summer.

We have shown that even if FoGSS forecasts have no skill, they will generally reflect climatology. An important exception to this may be in ephemeral catchments, for two main reasons:

1) Wapaba is designed for perennial catchments, and may not be able to replicate the highly non-linear and variable processes in ephemeral catchments;
2) Data censoring is only applied to observed flows for the maximum likelihood estimation of parameters, but not to simulated flows.

Wapaba is designed to produce low flows that approach, but never equal, zero. However, the conditional bias-correction applied at Stage 2 makes it possible for simulated flows to be zero. For catchments that only rarely have zero flows, the approach presented in this paper is adequate. However, it would be more theoretically correct to apply data censoring to both simulated and observed flows for catchments with many instances of zero flow in both observations and simulations. To do this would significantly complicate the parameter estimation procedure described in this paper. We will explore the performance of FoGSS in ephemeral catchments in future work. In the interim, we strongly caution against extrapolating the ability of FoGSS to perform at least as well as climatology, demonstrated in this paper, to forecasts for ephemeral rivers.

As we noted in Section 5.3, there is a general tendency of stochastic updating to induce positive bias at longer lead-times [*Wang et al.*, 2016]. This problem is not particularly acute in FoGSS because we only forecast to 12 lead-times. Nonetheless, it is it is a general weakness caused by treating each forecast as the median of the error distribution. In the transformed domain, the mean and median are approximately equivalent and the residual distribution is symmetrical. When data are back-transformed into the original domain the distribution is no longer symmetrical, meaning values less than the median are compressed relative to values greater than the median. This results in the mean taking larger values than the median. When this is repeated through multiple lead-times the mean can display bias (the median does not). The magnitude of the bias is related to the spread of the residual distribution, with larger residual variance resulting in larger bias. This is extremely difficult to avoid in systems that propagate hydrological uncertainty and which are trained only on simulations. The problem may be ameliorated through debiasing factors applied to the transformation [Guerrero, 1993] or by considering information from longer lead-times when building our model. This will be explored in future research.

It is feasible to include cryospheric processes in the FoGSS model. Snow accumulation/melt and glacier ablation models usually require forecasts of temperature (and often other variables). CBaM has been extended to temperature forecasts [*Schepen et al.*, 2016], and the inclusion of a module to handle cryospheric processes could be added to the Wapaba model. We note that forecasting in snow-free catchments is often considered more difficult than forecasting in cold regions, as the catchment memory imparted by cryospheric processes can be a significant contributor to forecast skill [e.g. *Mahanama et al.*, 2012].

Australia is somewhat unusual in already having two operational seasonal streamflow forecasting systems, but neither has been shown to replicate FoGSS' ability to produce reliable time series forecasts to long forecast horizons. The Bureau of Meteorology's (*the Bureau's*) seasonal streamflow forecasting service (www.bom.gov.au/water/ssf), based on the BJP, produces forecasts of total inflows for the coming 3-months. Recent work has shown that the BJP functions reasonably well in breaking down forecasts into 1-month streamflow [*Zhao et al.*, 2016], but the ability of the BJP to produce realistic time series to longer (>3-month) lead-times is untested. The Bureau also operates a pilot dynamic seasonal streamflow forecasting service [*Lerat et al.*, 2015; *Shin et al.*, 2011]. This makes use of POAMA forecasts downscaled with an analog method [*Charles et al.*, 2013] to force a daily rainfall runoff model to lead-times of 90 days. Hydrological uncertainty is characterized by sampling parameter uncertainty [*Kavetski et al.*, 2006]. To ensure forecasts are reliable, daily forecasts are aggregated to 1-month and 3-month totals and then post-processed [*Evin et al.*, 2014]. Unlike FoGSS, this final post-processing step lumps all uncertainties (rainfall and hydrological) into one error model, and it is applied only to lead-0 forecasts. The ability of the dynamic forecasting system to produce time series forecasts to multiple lead times is unknown.

Forecasts run at a daily time step may improve performance in catchments where catchment memory is shorter than one month, and they have obvious benefits for some application, e.g. where reservoir planning models run at a daily time step. However, daily forecasts are much noisier than monthly forecasts, making it generally more difficult to extract useful information at seasonal timescales. Maintaining the statistical properties of stochastic scenarios in long (>3-month) daily time series – e.g., ensuring that uncertainty is correctly represented and propagated, and that accumulations are reliable and take advantage of skill at early lead-times – is likely to be very challenging, essentially because statistical error models must operate over scores of lead-times, and are more likely to 'drift' away from realistic statistical properties. Accordingly, FoGSS may not be simply extended to daily forecasts without further development.

In principle, the FoGSS model could be operationalized fairly rapidly to provide forecast services. FoGSS has been implement in software developed and tested with the Bureau's requirements in mind. Generating cross-validated retrospective forecasts for ~30 years takes ~4 hours for each catchment, while generating a forecast takes only a few seconds. The Bureau's experience with operationalizing seasonal streamflow forecasts means they already have an extensive set of clear and standardized forecast outputs and verification measures which could be applied to Stage 3 FoGSS forecasts. We are currently working with the Bureau to operationalize the FoGSS error model.

7. SUMMARY AND CONCLUSIONS

We have developed a system for generating forecast guided stochastic scenarios (FoGSS) of streamflow to a forecast horizon of 12 months. FoGSS uses the method of calibration, bridging and merging (CBaM) of CGCM precipitation and sea surface temperature forecasts to produce ensemble precipitation forecasts that are reliable, unbiased at the catchment scale, and at skillful as possible. Rainfall forecasts are used to force the Wapaba rainfall-runoff model. Forecast uncertainty that arises from hydrological modelling is quantified and propagated by a 3-stage error model. The error model combines data transformation, bias-correction and a restricted autoregressive model to minimize forecast error and quantify the remaining forecast uncertainty. Hydrological forecast uncertainty is propagated through each 12-month forecast with a technique termed stochastic updating.

FoGSS makes use of skill from rainfall forecasts and catchment wetness to produce monthly time series forecasts that are as skillful as possible. Forecasts are typically skillful to lead-times of 2 or 3 months. As forecast skill declines with lead-time, the forecasts transit to stochastic scenarios. Each ensemble member is a plausible realization of streamflow for the coming 12 months. The statistical and temporal properties of the forecasts are realistic. Forecasts at individual months are reliable for all lead-times. This allows skill to propagate from short lead-times to forecasts of accumulated streamflow volumes: forecasts of 6-month and even 12-month streamflow volumes can be skillful in some instances.

The forecasts effectively combine the valuable information provided by a skillful forecast with the practicality of simple stochastic scenarios for the coming year. We believe FoGSS forecasts could be highly useful, and useable, by water agencies to improve water resources management.

8. ACKNOWLEDGEMENTS

This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Bureau of Meteorology and CSIRO Land & Water. Thanks to Senlin Zhou, Julien Lerat, Paul Feikema and Daehyok Shin (all Bureau of Meteorology) for supplying data and for fruitful discussions on the development of FoGSS. Thanks to Kelvin Michael (University of Tasmania) and Luis Neumann (CSIRO) for comments on an earlier draft of this manuscript. Catchment maps in Figure 1 are generated using data from the Bureau of Meteorology Geofabric (http://www.bom.gov.au/water/geofabric/). The streamflow and climate data used in this study are listed in the references or are available from the CSIRO data access portal http://data.csiro.au/dap/.

9. REFERENCES

Arnal, L., M. H. Ramos, E. Coughlan de Perez, H. L. Cloke, E. Stephens, F. Wetterhall, S. J. van Andel, and F. Pappenberger (2016), Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game, *Hydrol. Earth Syst. Sci.*, *20*(8), 3109-3128, doi: 10.5194/hess-20-3109-2016.

Ashok, K., H. Nakamura, and T. Yamagata (2007), Impacts of ENSO and Indian Ocean dipole events on the southern hemisphere storm-track activity during austral winter, *Journal of Climate*, *20*(13), 3147-3163, doi: 10.1175/jcli4155.1.

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt (2012), Skill of Real-Time Seasonal ENSO Model Predictions During 2002–11: Is Our Capability Increasing?, *Bulletin of the American Meteorological Society*, *93*(5), ES48-ES50, doi: 10.1175/bams-d-11-00111.2.

Bates, B. C., and E. P. Campbell (2001), A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, *37*(4), 937-947, doi: 10.1029/2000wr900363.

Bennett, J. C., Q. J. Wang, P. Pokhrel, and D. E. Robertson (2014), The challenge of forecasting high streamflows 1-3 months in advance with lagged climate indices in southeast Australia, *Natural Hazards and Earth System Sciences*, *14*(2), 219-233, doi: 10.5194/nhess-14-219-2014.

Bracken, C., B. Rajagopalan, and J. Prairie (2010), A multisite seasonal ensemble streamflow forecasting technique, *Water Resources Research*, *46*(3), W03532, doi: 10.1029/2009wr007965.

Bradley, A. A., M. Habib, and S. S. Schwartz (2015), Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach, *Water Resources Research*, *51*(9), 7382-7400, doi: 10.1002/2014wr016811.

Charles, A., B. Timbal, E. Fernandez, and H. Hendon (2013), Analog Downscaling of Seasonal Rainfall Forecasts in the Murray Darling Basin, *Monthly Weather Review*, *141*(3), 1099-1117, doi: 10.1175/mwr-d-12-00098.1.

Clark, M. P., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby (2004), The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields, *Journal of Hydrometeorology*, *5*, 243–262, doi: 10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Cox, D. R., and N. Reid (2004), A note on pseudolikelihood constructed from marginal densities, *Biometrika*, *91*(3), 729-737, doi: 10.1093/biomet/91.3.729.

Crochemore, L., M. H. Ramos, and F. Pappenberger (2016), Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, *2016*, 1-32, doi: 10.5194/hess-2016-78.

Crochemore, L., M.-H. Ramos, F. Pappenberger, S. J. van Andel, and A. W. Wood (2015), An experiment on risk-based decision-making in water management using monthly probabilistic forecasts, *Bulletin of the American Meteorological Society*, doi: 10.1175/bams-d-14-00270.1.

Day, G. N. (1985), Extended streamflow forecasting using NWSRFS, *Journal of Water Resources Planning and Management*, *111*(2), 157–170, doi: 10.1061/(ASCE)0733-9496(1985)111:2(157).

Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013), Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrology and Earth System Sciences*, *17*, 4209-4225, doi: 10.5194/hess-17-4209-2013.

Duan, Q., S. Sorooshian, and V. K. Gupta (1994), Optimal use of the SCE-UA global optimization method for calibrating watershed models, *Journal of Hydrology*, *158*(3–4), 265-284, doi: 10.1016/0022-1694(94)90057-4.
Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water Resources Research*, *49*, 4518-4524, doi: 10.1002/wrcr.20284.
Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resources Research*, *50*(3), 2350-2375, doi: 10.1002/2013wr014185.
Gneiting, T., and M. Katzfuss (2014), Probabilistic Forecasting, *Annual Review of Statistics and Its Application*, *1*(1), 125-151, doi: 10.1146/annurev-statistics-062713-085831.
Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona (2005), A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts, *Water Resources Research*, *41*(10), n/a-n/a, doi: 10.1029/2004wr003467.
Hawthorne, S., Q. J. Wang, A. Schepen, and D. E. Robertson (2013), Effective use of GCM outputs for forecasting monthly rainfalls to long lead times, *Water Resources Research*, *49*(9), 5427–5436, doi: 10.1002/wrcr.20453.
Hudson, D., A. G. Marshall, Y. Yin, O. Alves, and H. H. Hendon (2013), Improving Intraseasonal Prediction with a New Ensemble Generation Strategy, *Monthly Weather Review*, *141*(12), 4429-4449, doi: 10.1175/mwr-d-13-00059.1.
Kavetski, D., G. Kuczera, and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*(3), W03407, doi: 10.1029/2005wr004368.
Koster, R. D., S. P. P., Mahanama, B. Livneh, D. P. Lettenmaier, and R. H. Reichle (2010), Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow, *Nature Geoscience*, *3*, 613-616, doi: 10.1038/NGEO944.
Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology*, *331*(1–2), 161-177, doi: 10.1016/j.jhydrol.2006.05.010.
Lerat, J., et al. (2015), Dynamic streamflow forecasts within an uncertainty framework for 100 catchments in Australia, paper presented at HWRS 2015 Hydrology and Water Resources Symposium, Engineers Australia, Hobart.
Li, M., Q. J. Wang, and J. Bennett (2013), Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, *Water Resources Research*, *49*, 5913–5929, doi: 10.1002/wrcr.20445.
Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2015), A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, *Hydrol. Earth Syst. Sci.*, *19*(1), 1-15, doi: 10.5194/hess-19-1-2015.
Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2016), Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, *Hydrol. Earth Syst. Sci.*, *20*(9), 3561-3579, doi: 10.5194/hess-20-3561-2016.
Mahanama, S., B. Livneh, R. Koster, D. Lettenmaier, and R. Reichle (2012), Soil moisture, snow, and seasonal streamflow forecasts in the United States, *Journal of Hydrometeorology*, *13*(1), 189-203, doi: 10.1175/jhm-d-11-046.1.
Marshall, A. G., D. Hudson, M. C. Wheeler, O. Alves, H. H. Hendon, M. J. Pook, and J. S. Risbey (2014), Intra-seasonal drivers of extreme heat over Australia in observations

and POAMA-2, *Climate Dynamics*, *43*(7), 1915-1937, doi: 10.1007/s00382-013-2016-1.

Murphy, A. H. (1977), The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation, *Monthly Weather Review*, *105*(7), 803-816, doi: 10.1175/1520-0493(1977)105<0803:tvocca>2.0.co;2.

Najafi, M. R., H. Moradkhani, and T. C. Piechota (2012), Ensemble Streamflow Prediction: Climate signal weighting methods vs. Climate Forecast System Reanalysis, *Journal of Hydrology*, *442–443*, 105-116, doi: 10.1016/j.jhydrol.2012.04.003.

Olsson, J., C. B. Uvo, K. Foster, and W. Yang (2016), Technical Note: Initial assessment of a multi-method approach to spring-flood forecasting in Sweden, *Hydrol. Earth Syst. Sci.*, *20*(2), 659-667, doi: 10.5194/hess-20-659-2016.

Peng, Z., Q. J. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. Wang (2014), Statistical Calibration and Bridging of ECMWF System4 Outputs for Forecasting Seasonal Precipitation over China *Journal of Geophysical Research (Atmospheres)*, *119*, 7116–7135, doi: 10.1002/2013JD021162.

Priestley, C. H. B., and R. J. Taylor (1972), On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, *Monthly Weather Review*, *100*(2), 81-92, doi: 10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2.

Ramos, M. H., S. J. van Andel, and F. Pappenberger (2013), Do probabilistic forecasts lead to better decisions?, *Hydrology and Earth System Sciences*, *17*(6), 2219-2232, doi: 10.5194/hess-17-2219-2013.

Raupach, M. R., P. R. Briggs, V. Haverd, E. A. King, M. Paget, and C. M. Trudinger (2008), Australian Water Availability Project (AWAP), final report for Phase 3*Rep.*, 67 pp pp, CSIRO Marine and Atmospheric Research, Canberra, Australia.

Regonda, S. K., B. Rajagopalan, and M. Clark (2006), A new method to produce categorical streamflow forecasts, *Water Resources Research*, *42*(9), n/a-n/a, doi: 10.1029/2006wr004984.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, *46*(5), W05521, doi: 10.1029/2009wr008328.

Robertson, D. E., P. Pokhrel, and Q. J. Wang (2013), Improving statistical forecasts of seasonal streamflows using hydrological model output, *Hydrology and Earth System Sciences*, *17*(2), 579-593, doi: 10.5194/hess-17-579-2013.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata (1999), A dipole mode in the tropical Indian Ocean, *Nature*, *401*(6751), 360-363, doi: 10.1038/43854.

Schaefli, B., D. B. Talamba, and A. Musy (2007), Quantifying hydrological modeling errors through a mixture of normal distributions, *Journal of Hydrology*, *332*, 303-315, doi: 10.1016/j.jhydrol.2006.07.005.

Schepen, A., and Q. J. Wang (2013), Toward accurate and reliable forecasts of australian seasonal rainfall by calibrating and merging multiple coupled GCMS, *Monthly Weather Review*, *141*(12), 4554-4563, doi: 10.1175/mwr-d-12-00253.1.

Schepen, A., and Q. Wang (2014), Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output, *Journal of Hydrology*, *519*, 2920–2931, doi: 10.1016/j.jhydrol.2014.03.017.

Schepen, A., and Q. J. Wang (2015), Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia, *Water Resources Research*, *51*(3), 1797-1812, doi: 10.1002/2014wr016163.

Schepen, A., Q. J. Wang, and D. E. Robertson (2012), Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall, *Journal of Geophysical Research*, *117*, D20107, doi: 10.1029/2012JD018011.
Schepen, A., Q. J. Wang, and D. E. Robertson (2014), Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs, *Monthly Weather Review*, *142*(5), 1758-1770, doi: 10.1175/mwr-d-13-00248.1.
Schepen, A., Q. J. Wang, and Y. Everingham (2016), Calibration, bridging and merging to improve GCM seasonal temperature forecasts in Australia, *Monthly Weather Review*, *144*(6), 2421-2441, doi: 10.1175/MWR-D-15-0384.1.
Schepen, A., Q. J. Wang, and D. E. Robertson (2018), Application to post-processing of meteorological seasonal forecasting, in *Handbook of hydrometeorological ensemble forecasting*, edited by Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke and J. C. Schaake, Springer-Verlag Berlin Heidelberg, doi: 10.1007/978-3-642-40457-3_18-1.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resources Research*, *46*(10), W10531, doi: 10.1029/2009wr008933.
Seo, D. J., H. D. Herr, and J. C. Schaake (2006), A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, *Hydrology and Earth System Sciences Discussions*, *3*, 1987-2035, doi: 10.5194/hessd-3-1987-2006.
Shin, D., A. Schepen, T. Peatey, S. Zhou, A. MacDonald, T. Chia, J. Perkins, and N. Plummer (2011), WAFARi: A new modelling system for Seasonal Streamflow Forecasting service of the Bureau of Meteorology, Australia, paper presented at 19th International Congress on Modelling and Simulation, MSSANZ, Perth, Australia, 12-16 December 2011.
Shukla, S., and D. P. Lettenmaier (2011), Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, *Hydrology and Earth System Sciences*, *15*, 3529–3538, doi: 10.5194/hess-15-3529-2011.
Smith, T., L. Marshall, and A. Sharma (2015), Modeling residual hydrologic errors with Bayesian inference, *Journal of Hydrology*, *528*, 29-37, doi: 10.1016/j.jhydrol.2015.05.051.
Svensson, C., A. Brookshaw, A. A. Scaife, V. A. Bell, J. D. Mackay, C. R. Jackson, J. Hannaford, H. N. Davies, A. Arribas, and S. Stanley (2015), Long-range forecasts of UK winter hydrology, *Environmental Research Letters*, *10*(6), 064006, doi: 10.1088/1748-9326/10/6/064006.
van Dijk, A. I. J. M., H. E. Beck, R. S. Crosbie, R. A. M. de Jeu, Y. Y. Liu, G. M. Podger, B. Timbal, and N. R. Viney (2013), The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society, *Water Resources Research*, *49*, 1-18, doi: 10.1002/wrcr.20123.
Verdon, D. C., and S. W. Franks (2005), Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia, *Water Resources Research*, *41*, W09413, doi: 10.1029/2004WR003845.
Wang, Q. J., and D. E. Robertson (2011), Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, *Water Resources Research*, *47*, W02546, doi: 10.1029/2010WR009333.

Wang, Q. J., D. E. Robertson, and F. H. S. Chiew (2009), A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resources Research*, *45*, W05407, doi: 10.1029/2008WR007355.

Wang, Q. J., A. Schepen, and D. E. Robertson (2012a), Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging, *Journal of Climate*, *25*, 5524-5537, doi: 10.1175/JCLI-D-11-00386.1.

Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel (2012b), A log-sinh transformation for data normalization and variance stabilization, *Water Resources Research*, *48*, W05514, doi: 10.1029/2011WR010973.

Wang, Q. J., M. Li, J. C. Bennett, and D. E. Robertson (2016), Stochastic updating for ensemble hydrological forecasting, in *Water Infrastructure and the Environment: 56th New Zealand hydrological society and the 37th Australian hydrological and water resources symposium including the 7th IPENZ rivers group*, edited by C. Pearson and M. Babister, Engineers Australia, Queenstown, New Zealand.

Wang, Q. J., T. C. Pagano, S. L. Zhou, H. A. P. Hapuarachchi, L. Zhang, and D. E. Robertson (2011), Monthly versus daily water balance models in simulating monthly runoff, *Journal of Hydrology*, *404*(3–4), 166-175, doi: 10.1016/j.jhydrol.2011.04.027.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay (2004), Climate Index Weighting Schemes for NWS ESP-Based Seasonal Volume Forecasts, *Journal of Hydrometeorology*, *5*(6), 1076-1090, doi: 10.1175/jhm-381.1.

Wood, A. W., and J. C. Schaake (2008), Correcting Errors in Streamflow Forecast Ensemble Mean and Spread, *Journal of Hydrometeorology*, *9*(1), 132-148, doi: 10.1175/2007jhm862.1.

Wood, A. W., and D. P. Lettenmaier (2008), An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophysical Research Letters*, *35*(14), L14401, doi: 10.1029/2008gl034648.

Yeo, I. K., and R. A. Johnson (2000), A new family of power transformations to improve normality or symmetry, *Biometrika*, *87*(4), 954–959, doi: 10.1093/biomet/87.4.954.

Yuan, X. (2016), An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 2: The added value from climate forecast models, *Hydrol. Earth Syst. Sci.*, *20*(6), 2453-2466, doi: 10.5194/hess-20-2453-2016.

Yuan, X., E. F. Wood, and Z. Ma (2015a), A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, *Wiley Interdisciplinary Reviews: Water*, *2*(5), 523-536, doi: 10.1002/wat2.1088.

Yuan, X., J. K. Roundy, E. F. Wood, and J. Sheffield (2015b), Seasonal Forecasting of Global Hydrologic Extremes: System Development and Evaluation over GEWEX Basins, *Bulletin of the American Meteorological Society*, *96*(11), 1895-1912, doi: 10.1175/bams-d-14-00003.1.

Zappa, M., S. Jaun, U. Germann, A. Walser, and F. Fundel (2011), Superposition of three sources of uncertainties in operational flood forecasting chains, *Atmospheric Research*, *100*(2–3), 246-262, doi: 10.1016/j.atmosres.2010.12.005.

Zhao, T., A. Schepen, and Q. J. Wang (2016), Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach, *Journal of Hydrology*, *published online*, doi: 10.1016/j.jhydrol.2016.07.040.

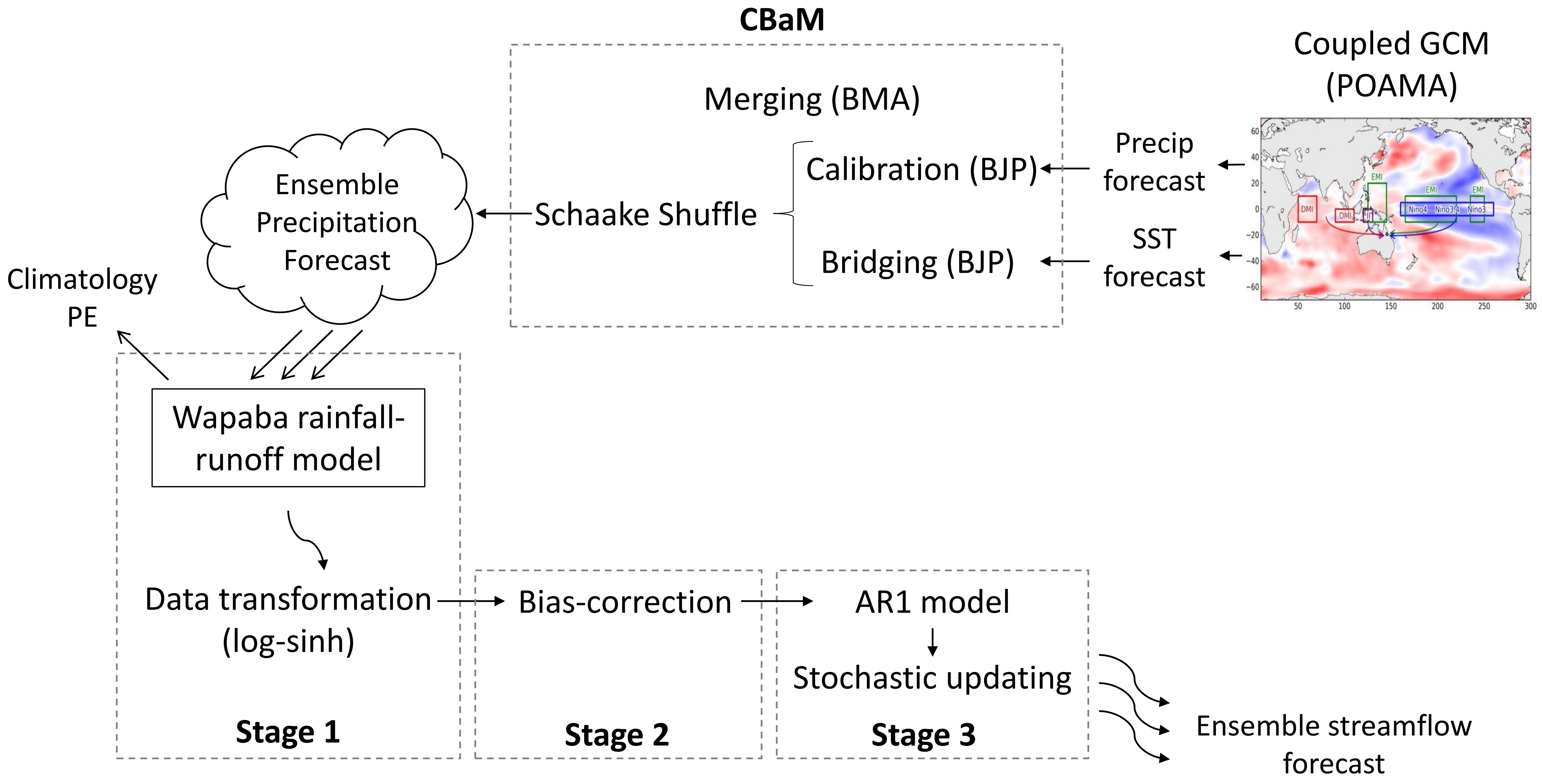
Appendix A

Example Wapaba parameters estimated at Stage 1 without cross-validation (i.e. using all data from the period 1982-2009) are given in Table A1. See Section 2.2 for details of the parameter estimation procedure.

Table A1 Example Wapaba and error model parameter sets estimated using all data for 1982-2009

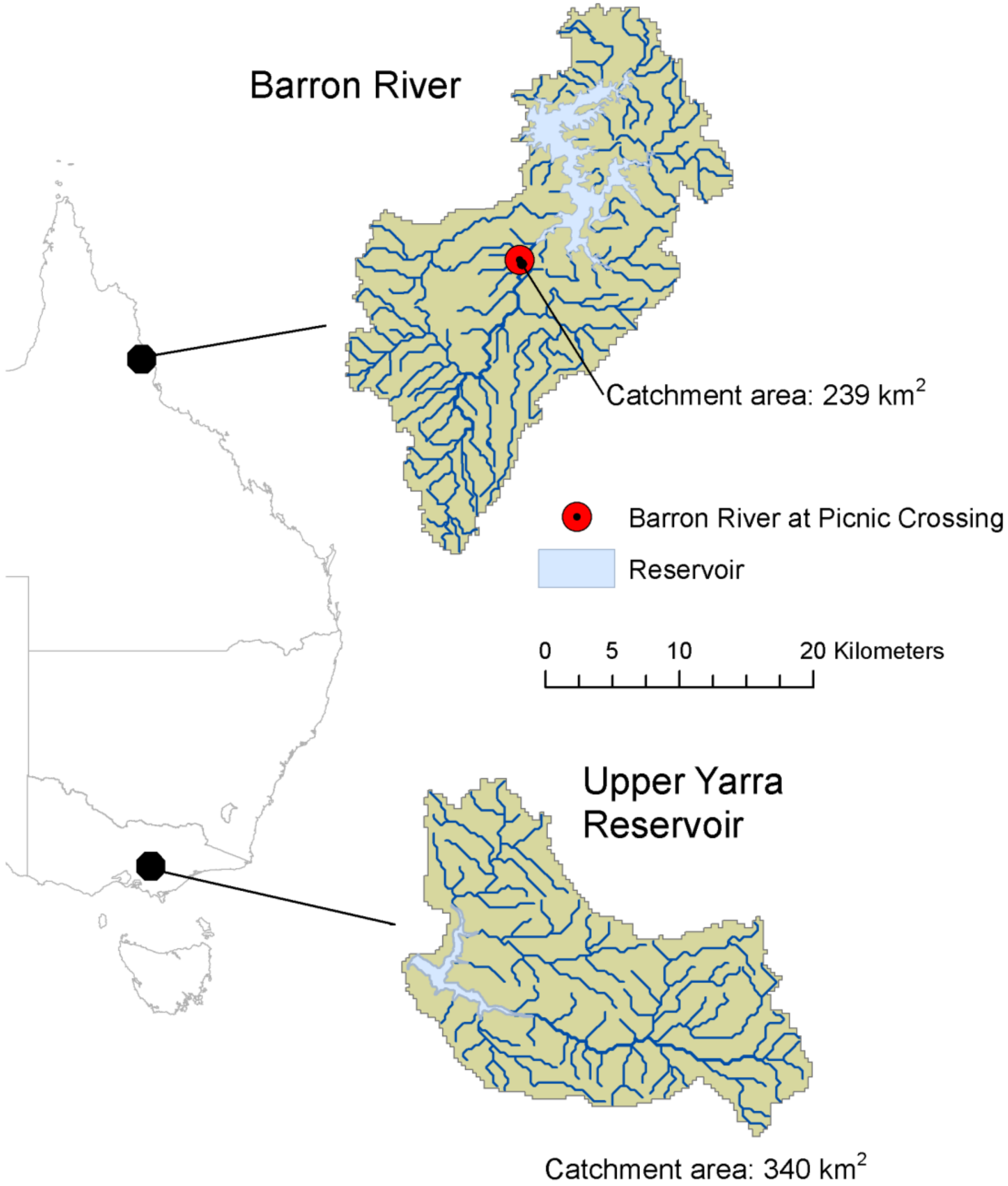
Parameters	*Barron River*	*Yarra River*
α_1	2.11	2.08
α_2	2.06	0.99
β	0.68	0.72
k^{-1}	0.014	0.04
$S_{\max}$	370	559

Figure 1. Figure



CBaM
Coupled GCM (POAMA)
Merging (BMA)
Calibration (BJP)
Precip forecast
Schaake Shuffle
Bridging (BJP)
SST forecast
Ensemble Precipitation Forecast
Climatology PE
Wapaba rainfall-runoff model
Data transformation (log-sinh)
Bias-correction
AR1 model
Stochastic updating
Stage 1
Stage 2
Stage 3
Ensemble streamflow forecast

Figure 2. Figure


Barron River
Catchment area: 239 km²
Barron River at Picnic Crossing
Reservoir
0 5 10 20 Kilometers
Upper Yarra Reservoir
Catchment area: 340 km²

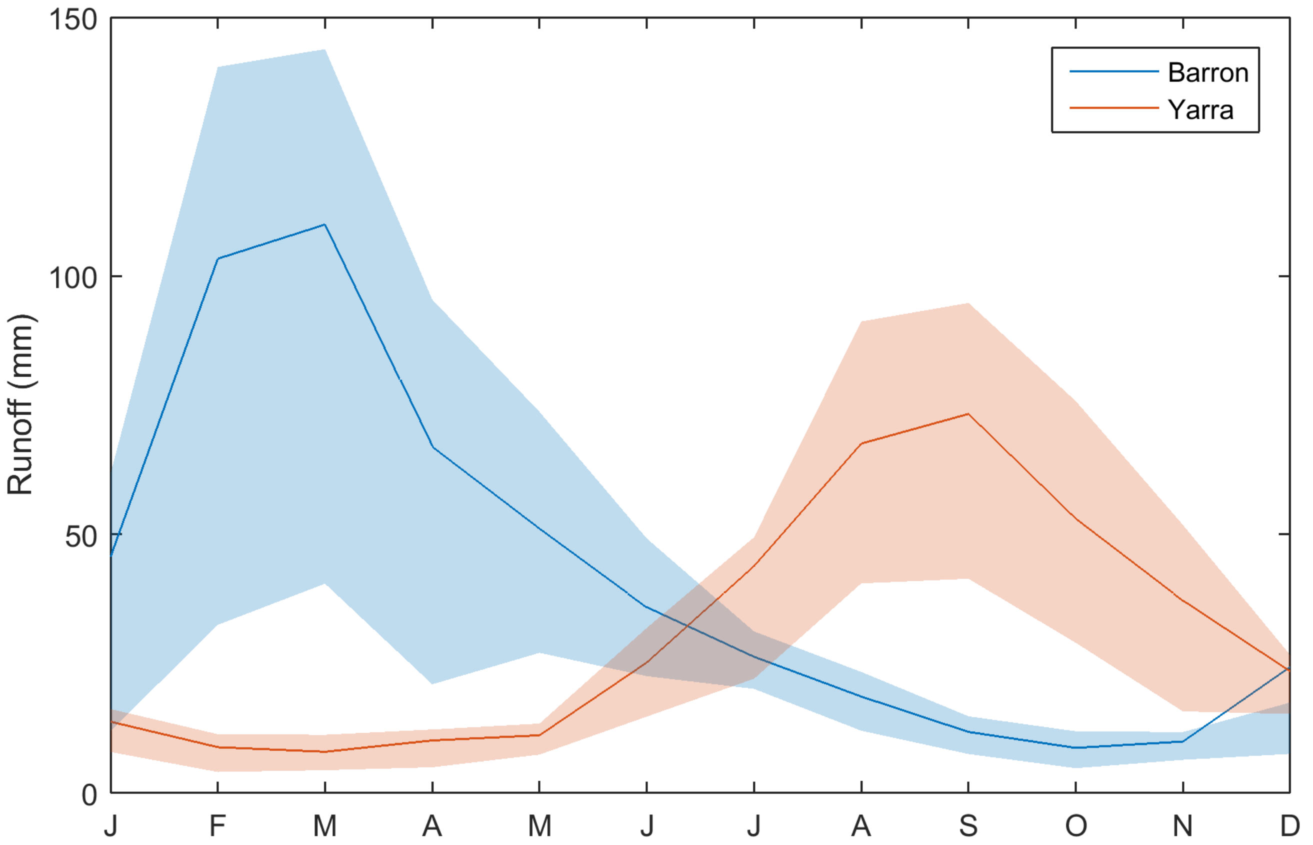

Runoff (mm)
150
100
50
0
J F M A M J J A S O N D
Barron
Yarra

Figure 3. Figure

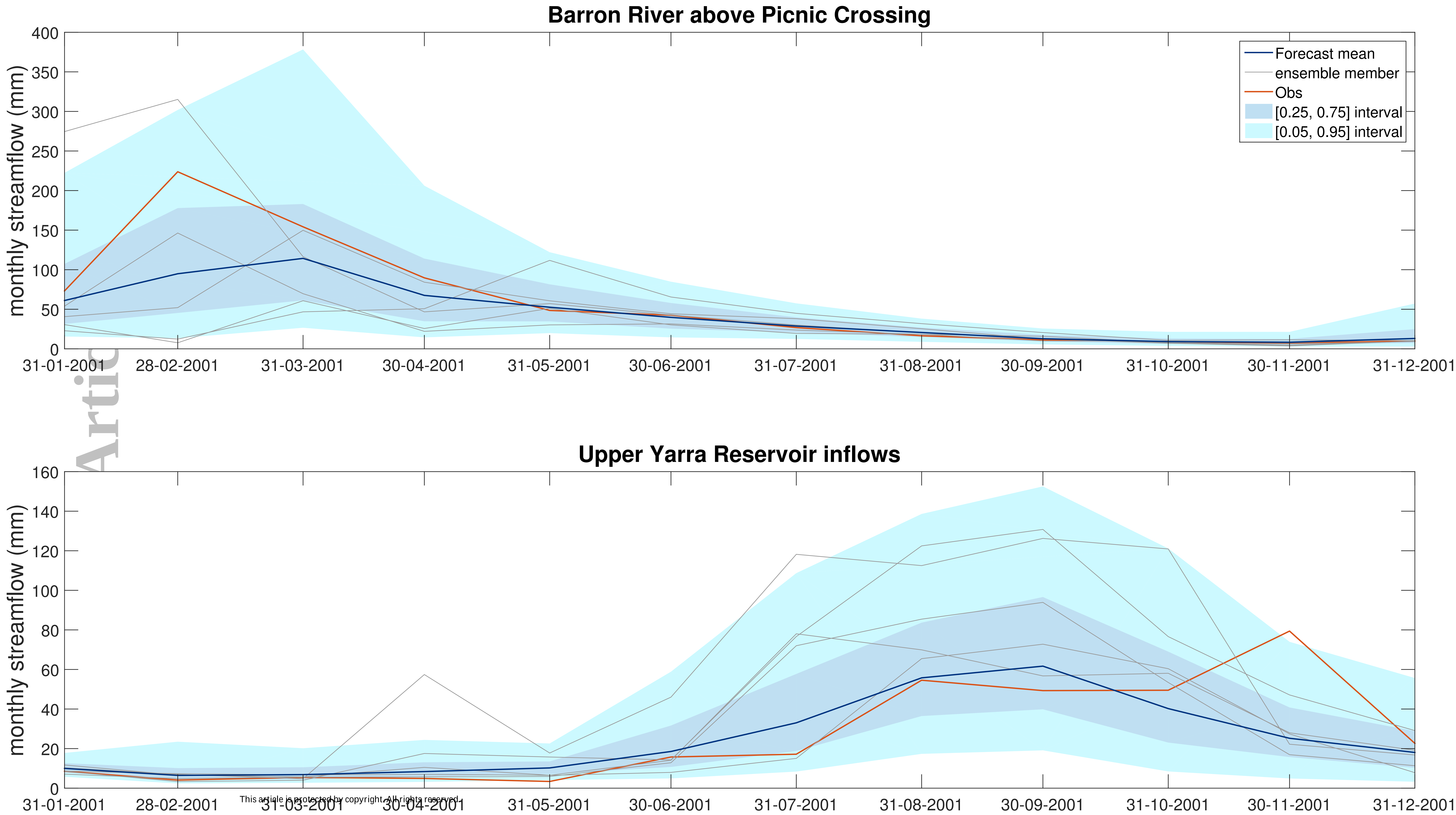
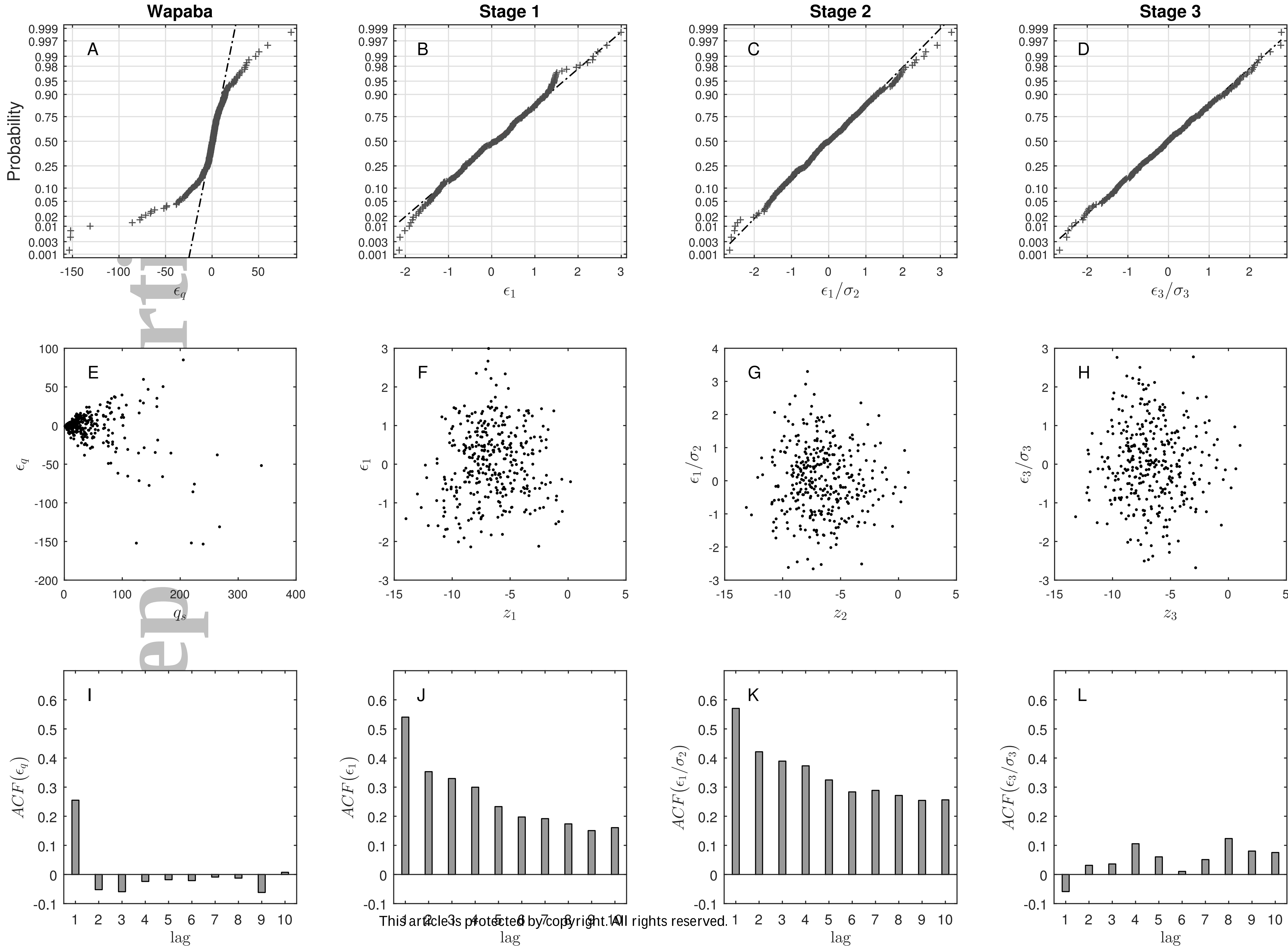
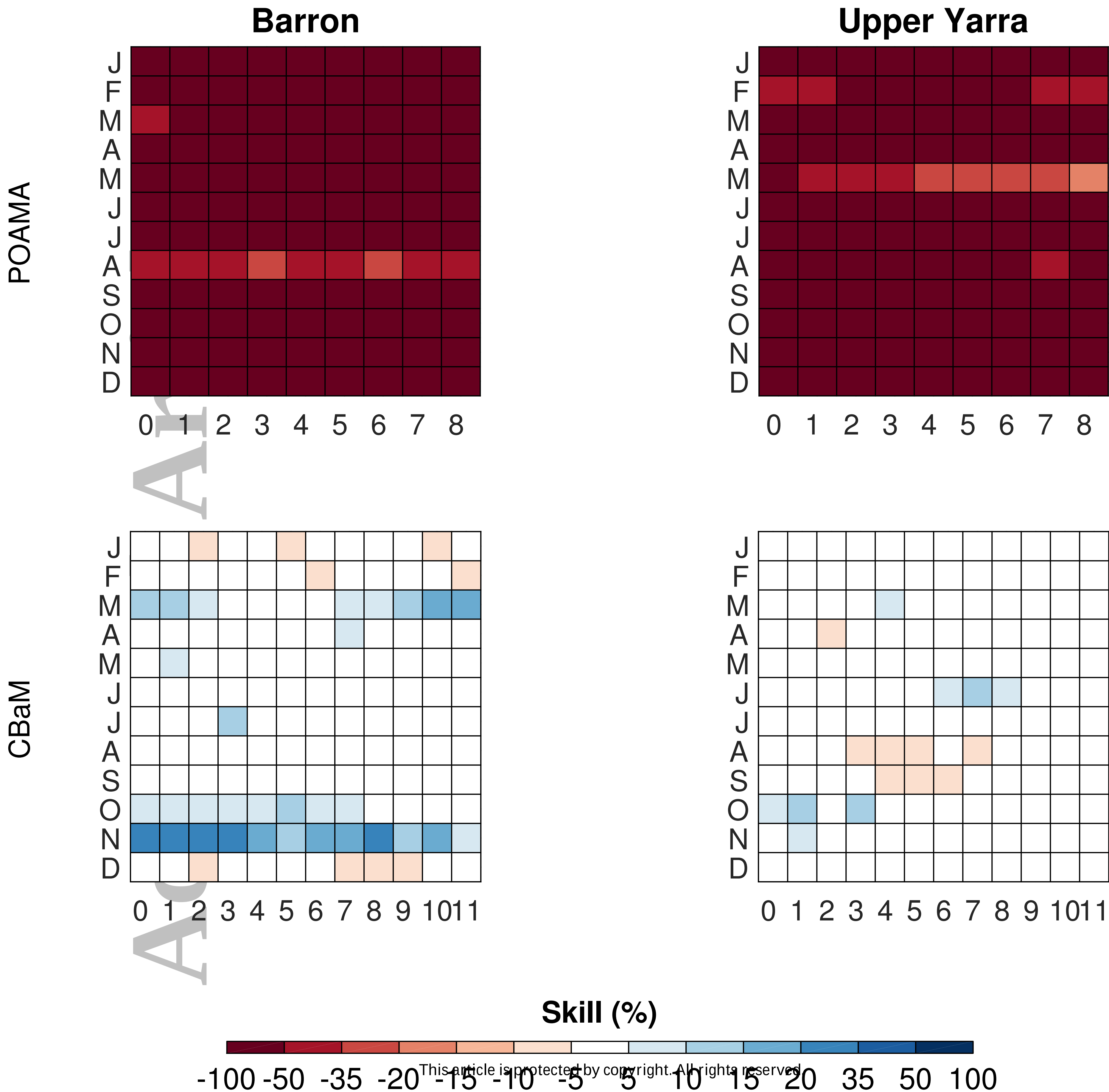

Barron River above Picnic Crossing
monthly streamflow (mm)
Forecast mean
ensemble member
Obs
[0.25, 0.75] interval
[0.05, 0.95] interval
31-01-2001
28-02-2001
31-03-2001
30-04-2001
31-05-2001
30-06-2001
31-07-2001
31-08-2001
30-09-2001
31-10-2001
30-11-2001
31-12-2001
Upper Yarra Reservoir inflows
monthly streamflow (mm)

Figure 4. Figure



Wapaba
Stage 1
Stage 2
Stage 3
A
B
C
D
E
F
G
H
I
J
K
L
Probability
ϵ_q
ϵ_1
ϵ_1/σ_2
ϵ_3/σ_3
q_s
z_1
z_2
z_3
$ACF(\epsilon_q)$
$ACF(\epsilon_1)$
$ACF(\epsilon_1/\sigma_2)$
$ACF(\epsilon_3/\sigma_3)$
lag

Figure 5. Figure



Barron
Upper Yarra
POAMA
CBaM
J
F
M
A
M
J
J
A
S
O
N
D
0 1 2 3 4 5 6 7 8
0 1 2 3 4 5 6 7 8 9 10 11
Skill (%)
-100 -50 -35 -20 -15 -10 -5 5 10 15 20 35 50 100

Figure 6. Figure

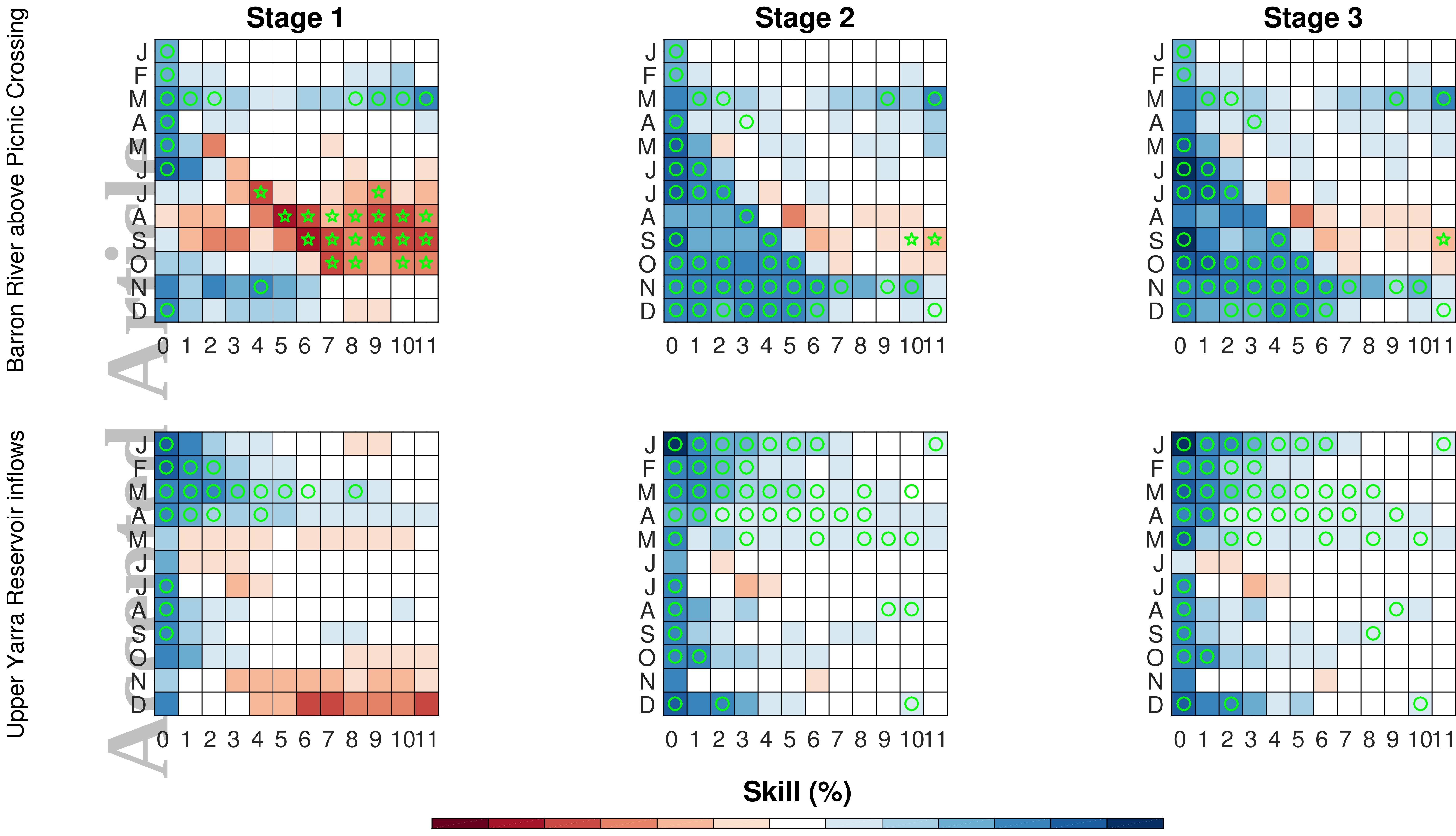
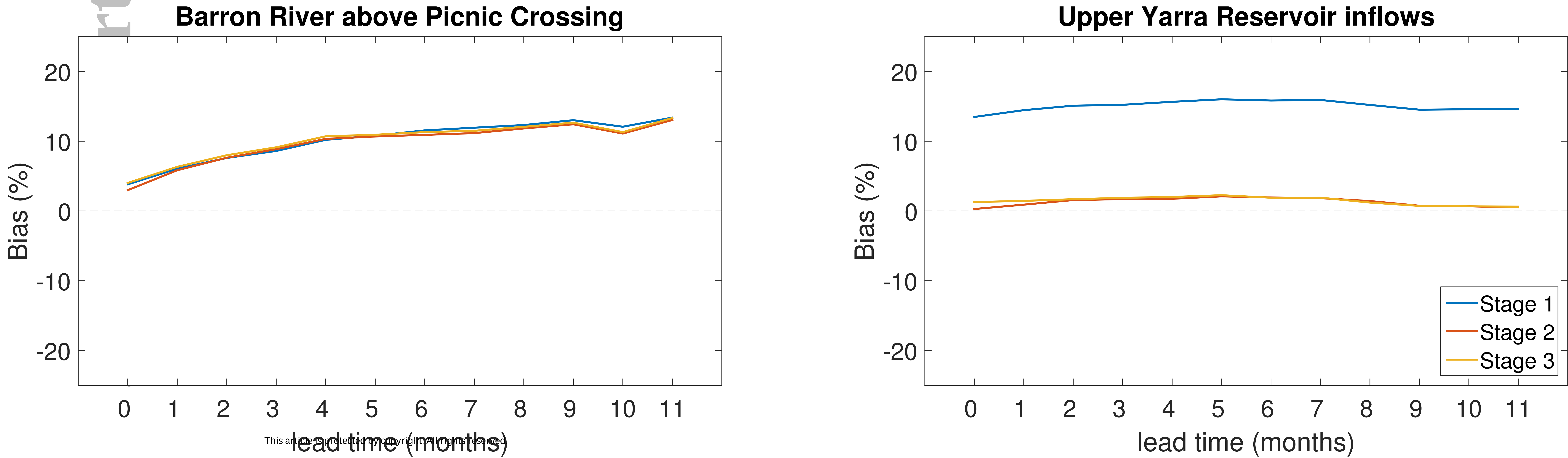

Stage 1
Stage 2
Stage 3
Barron River above Picnic Crossing
Upper Yarra Reservoir inflows
J
F
M
A
M
J
J
A
S
O
N
D
0 1 2 3 4 5 6 7 8 9 10 11
Skill (%)
-100 -50 -35 -20 -15 -10 -5 5 10 15 20 35 50 100

Figure 7. Figure



Barron River above Picnic Crossing
Upper Yarra Reservoir inflows
Bias (%)
20
10
0
-10
-20
0
1
2
3
4
5
6
7
8
9
10
11
lead time (months)
Stage 1
Stage 2
Stage 3

Figure 8. Figure

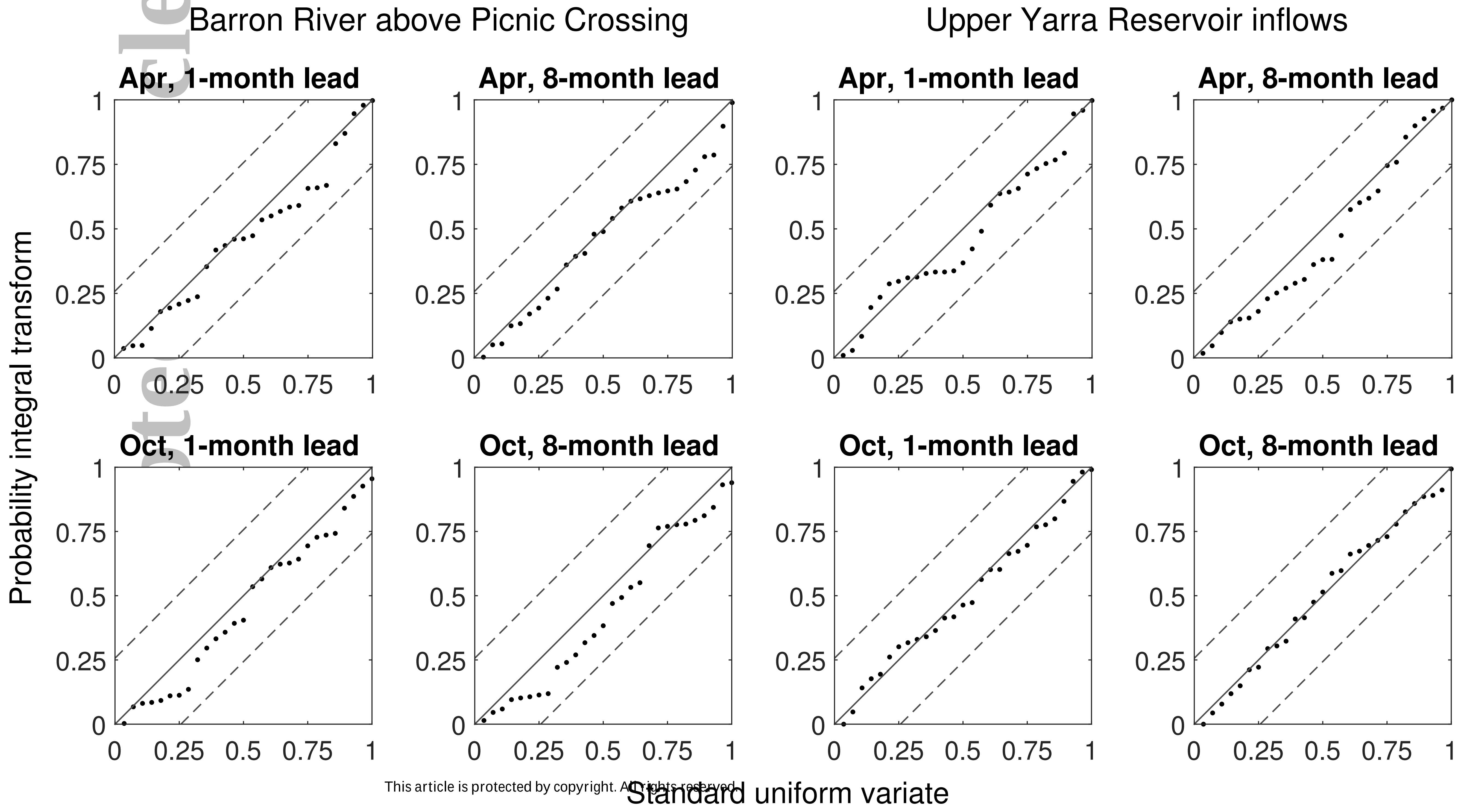
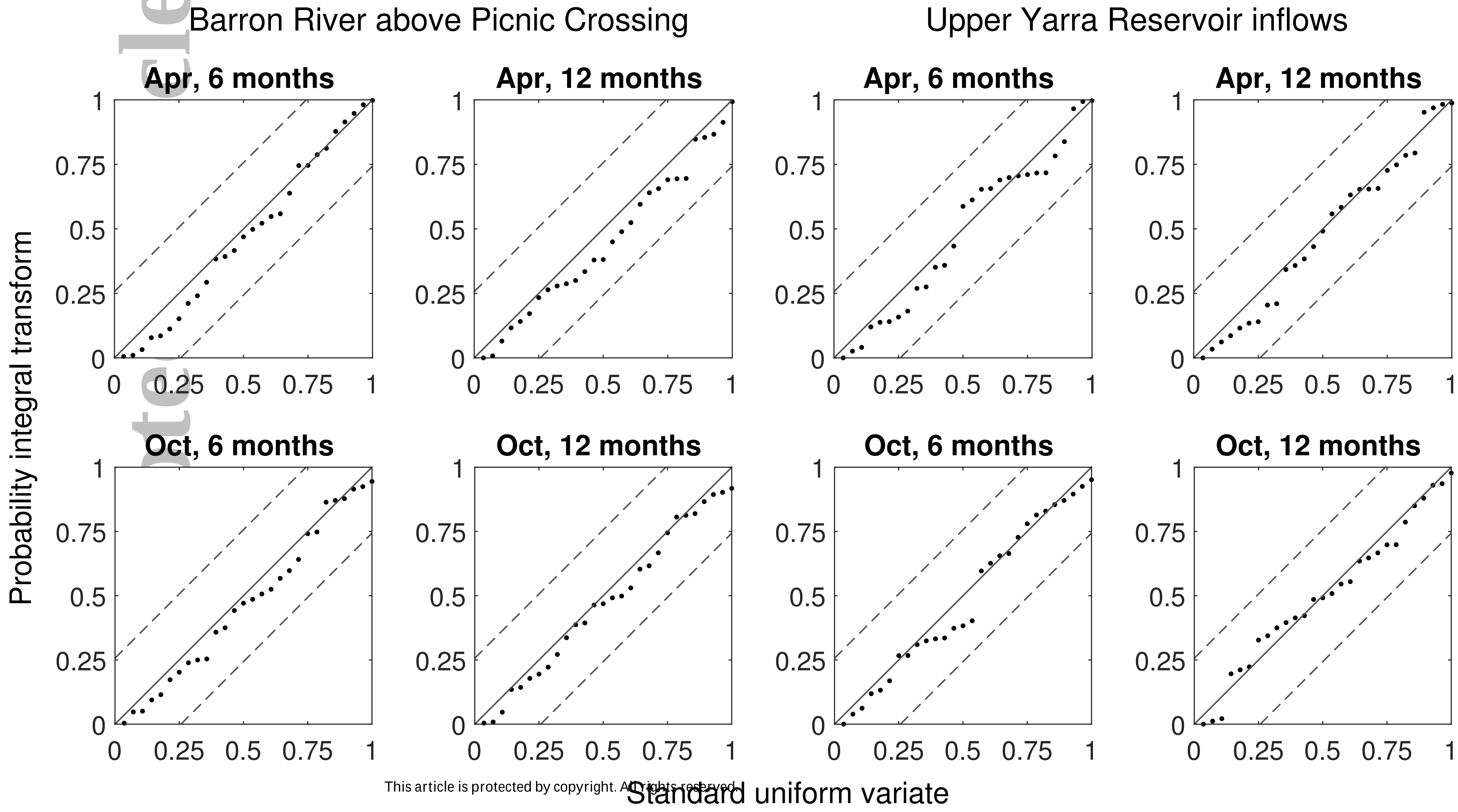
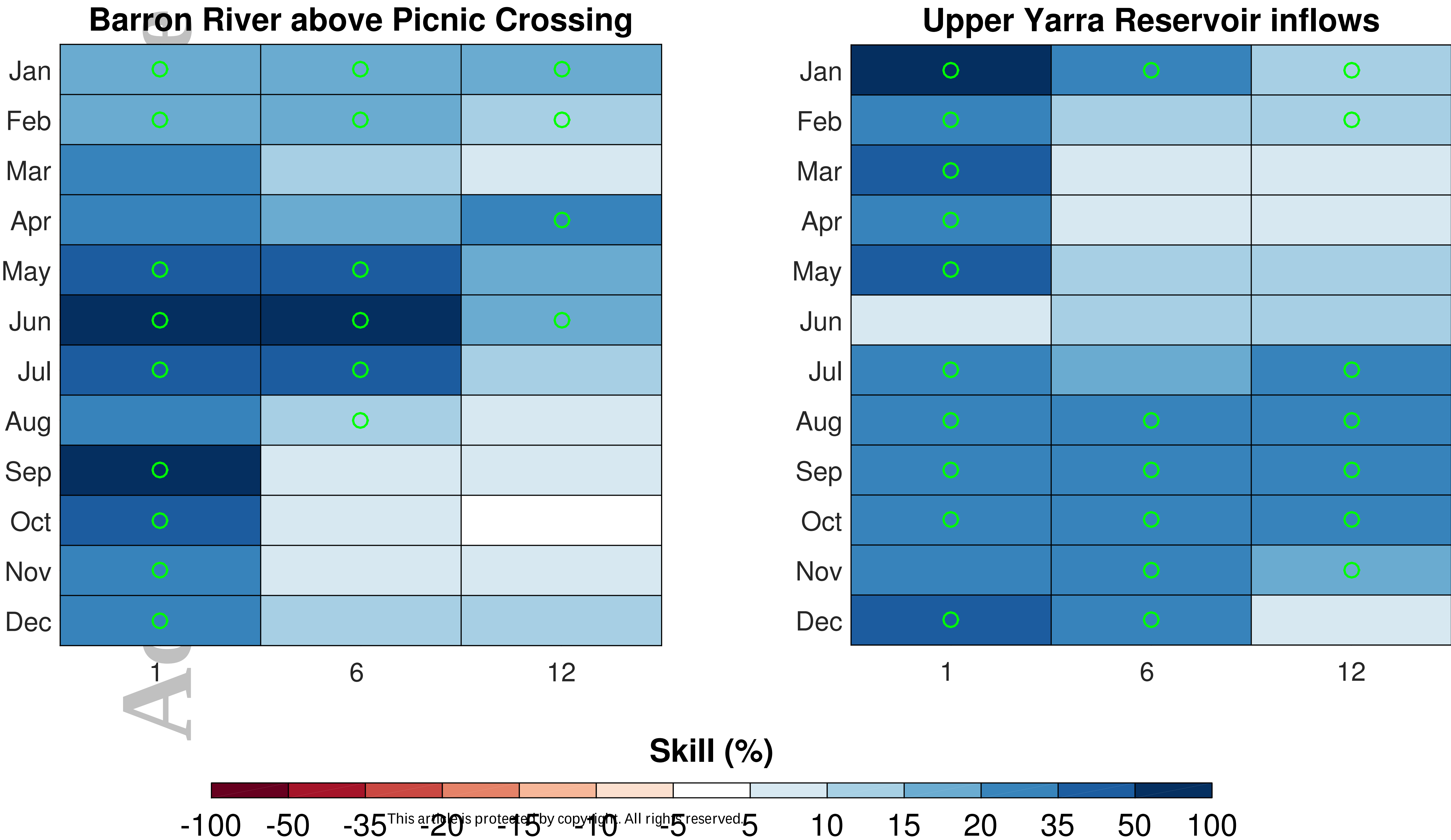

Barron River above Picnic Crossing
Upper Yarra Reservoir inflows
Apr, 1-month lead
Apr, 8-month lead
Apr, 1-month lead
Apr, 8-month lead
Oct, 1-month lead
Oct, 8-month lead
Oct, 1-month lead
Oct, 8-month lead
Probability integral transform
Standard uniform variate
0
0.25
0.5
0.75
1

Figure 9. Figure



Barron River above Picnic Crossing
Upper Yarra Reservoir inflows
Apr, 6 months
Apr, 12 months
Apr, 6 months
Apr, 12 months
Oct, 6 months
Oct, 12 months
Oct, 6 months
Oct, 12 months
Probability integral transform
Standard uniform variate
0
0.25
0.5
0.75
1

Figure 10. Figure


Barron River above Picnic Crossing
Upper Yarra Reservoir inflows
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
1
6
12
Skill (%)
-100 -50 -35 -20 -15 -10 -5 5 10 15 20 35 50 100