Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting

Ming Li[1], Q.J. Wang[2], James C. Bennett[2] and David E. Robertson[2]
[1]CSIRO Data61, Floreat, WA, Australia
[2]CSIRO Land and Water, Clayton, Victoria, Australia

Corresponding Author:
Dr Ming Li
CSIRO Data61
Private Bag 5, Wembley, WA 6014
Australia
Phone +61-8-9333 6417
Fax +61-8-9333 6121
Email Ming.Li@csiro.au

ABSTRACT:

This study develops a new error modelling method for ensemble short-term and real-time streamflow forecasting, called error reduction and representation in stages (ERRIS). The novelty of ERRIS is that it does not rely on a single complex error model but runs a sequence of simple error models through four stages. At each stage, an error model attempts to incrementally improve over the previous stage. Stage 1 establishes parameters of a hydrological model and parameters of a transformation function for data normalization, Stage 2 applies a bias-correction, Stage 3 applies autoregressive (AR) updating, and Stage 4 applies a Gaussian mixture distribution to represent model residuals. In a case study, we apply ERRIS for one-step ahead forecasting at a range of catchments. The forecasts at the end of Stage 4 are shown to be much more accurate than at Stage 1 and to be highly reliable in representing forecast uncertainty. Specifically, the forecasts become more accurate by applying the AR updating at Stage 3, and more reliable in uncertainty spread by using a mixture of two Gaussian distributions to represent the residuals at Stage 4. ERRIS can be applied to any existing calibrated hydrological models, including those calibrated to deterministic (e.g. least-squares) objectives.

KEYWORDS: streamflow forecasting, updating, residual distribution, multi-stage error modelling, ensemble forecasting

1. Introduction

Streamflow forecasts have long been used to support management of river conditions, such as flood emergency response and optimal water allocation. Recently, much research has been carried out on ensemble streamflow forecasting [e.g. *Alfieri et al.*, 2013; *Bennett et al.*, 2014a; *Demargne et al.*, 2014; *Thielen et al.*, 2009], encouraged by research communities such as the Hydrological Ensemble Prediction Experiment (HEPEX - http://hepex.org/). In recognition that streamflow forecasts can be subject to significant errors, forecast ensembles are used to represent forecast uncertainty. In producing ensemble forecasts, one aims to reduce forecast uncertainty as much as possible to give the most accurate forecasts. One also aims to represent the remaining forecast uncertainty reliably to give the right distribution among ensemble members.

Streamflow forecasts are usually made by initializing hydrological models (e.g. conceptual rainfall-runoff models) and then forcing them with forecast rainfall. There are a number of sources of errors in streamflow forecasts, including errors in measurement of rainfall and streamflow, errors in hydrological model structure, errors in model parameters, and errors in forecast rainfall. Ideal hydrological error quantification would account for each individual source of errors explicitly and reliably, such that all sources of errors would accumulate to accurately represent overall errors in the streamflow forecasts. Various attempts have been made to identify and decompose the sources of errors, by methods such as sequential optimization and data assimilation [*Vrugt et al.*, 2005], sequential assimilation [*Moradkhani et al.*, 2005], the Bayesian total error analysis (BATEA) [*Kavetski et al.*, 2006a; b; *Kuczera et al.*, 2006], and Integrated Bayesian Uncertainty Estimator (IBUNE) [*Ajami et al.*, 2007]. Such methods are useful for attempting to separate the major sources of errors, identifying deficiencies of model structure,

67 performing parameter sensitivity analyses and comparing different hydrological models, without
68 confounding input and output errors. However, because of a lack of information on the different
69 sources of errors and on how they interact with each other, it is highly challenging to apply an
70 error decomposition approach to arrive at statistically reliable overall errors in streamflow
71 forecasts [*Renard et al.*, 2010].

72 An alternative approach is to consider only the overall errors of forecasts, without attempting to
73 explain the sources of errors. An estimate of the overall error of a forecast is the residual, defined
74 as the difference between modelled streamflow and observations. We now concentrate our
75 discussion on residuals, but we will continue to refer to models of residuals as 'error models',
76 following common practice. Residuals of a series of forecasts form a time series. The most
77 traditional and simplest error model, related to the classical least squares calibration, is based on
78 the assumption of uncorrelated homoscedastic Gaussian residuals in the time series of residuals
79 [*Diskin and Simon*, 1977]. This assumption is generally not valid for hydrological applications,
80 where residuals are frequently auto-correlated, heteroscedastic and non-Gaussian [*Kuczera*,
81 1983; *Sorooshian and Dracup*, 1980]. More sophisticated error models have been developed to
82 address correlation, variance structure and the distribution of residuals. Autoregressive models
83 have been widely used to account for auto-correlation of residuals [e.g. *Bates and Campbell*,
84 2001; *Xiong and O'Connor*, 2002]. Heteroscedasticity may be explicitly dealt with by describing
85 the variance of residuals as a function of some state-dependent variables (e.g. observed
86 streamflow, dry/wet seasons) [e.g. *Evin et al.*, 2013; *Pianosi and Raso*, 2012; *Schaefli et al.*,
87 2007; *Yang et al.*, 2007]. Non-Gaussianity of residuals may be explicitly represented by non-
88 Gaussian probability distributions [e.g. *Marshall et al.*, 2006; *Schaefli et al.*, 2007; *Schoups and*
89 *Vrugt*, 2010]. Heteroscedasticity and non-Gaussianity of residuals may also be dealt with

implicitly, and often more conveniently, by using data transformation to normalize the residuals and stabilize their variance, such as the normal quantile transform [*Kelly and Krzysztofowicz*, 1997; *Krzysztofowicz*, 1997; *Montanari and Brath*, 2004], the Box-Cox transformation [*Thyer et al.*, 2002] and the log-sinh transformation [*Wang et al.*, 2012]. *Solomatine and Shrestha* [2009] presented an alternative method of predicting residual error distributions using machine learning techniques. They built a non-linear regression model to predict the forecast quantiles at each lead time. Their method is not based on an autoregressive model but captured recent information about the model error with a non-linear regression.

Broadly, previous attempts to model residuals can be divided into 'post-processor' methods that separate the estimation of hydrological model parameters from the estimation of error model parameters, and 'joint inference' methods that estimate all parameters at once. Post-processor methods (e.g. *Evin et al.* [2014]] are often held to be less theoretically desirable than joint inference methods [e.g. *Kuczera*, 1983; *Bates and Campbell*, 2001]. This is because joint inference methods aspire to a complete description of the behavior of errors, including behaviors that arise from interactions between parameters from hydrological and error models [see discussion in *Evin et al.*, 2014]. Unfortunately joint inference methods can have serious limitations for operational forecasting of streamflows. *Li et al.* [2015] showed that a joint inference method caused poor performance in the hydrological model when it was isolated from the error model (we will call this the 'base' hydrological model). Error models that account for auto-correlated residuals usually have less influence on forecasts as lead-time increases. Thus as lead-time increases, and the influence of the error model decreases, the quality of the forecast relies on the performance of the base hydrological

model and the quality of meteorological forecasts [*Bennett et al.*, 2014a]. *Evin et al.* [2014] demonstrated another (and perhaps more egregious) limitation of joint inference methods: joint estimation can result in deleterious interference between error model and hydrological model parameters, leading to poor out-of-sample streamflow predictions. In our experience, interactions between parameters of the hydrological model and the error model can make it very difficult to calibrate the models jointly. The shape of the distribution of forecast residuals can change markedly after hydrological model forecasts are updated, for example with an autoregressive error model. Despite considerable progress in hydrological uncertainty modelling, few studies in the literature present model forecasts (or simulations) that are practically reliable when error updating is applied [e.g. *Gragne et al.*, 2015; *Schoups and Vrugt*, 2010].

This paper develops a new error modelling method, called error reduction and representation in stages (ERRIS), for real-time and short-term streamflow forecasting applications. ERRIS is a further development of the restricted autoregressive model [*Li et al.*, 2015] and a seasonal error model developed by *Li et al. [2013]*. ERRIS is a post-processing method developed to deal with the overall errors of streamflow forecasts resulting from hydrological uncertainty only. We assume that errors in streamflow forecasts due to weather forecasts (precipitation in particular) will be considered separately by using ensemble weather forecasts [*Bennett et al.*, 2014a; *Robertson et al.*, 2013; *Shrestha et al.*, 2013], and we do not consider these in this paper. For convenience, in this study we use the term *streamflow forecast* to mean one-step-ahead model prediction of streamflow, given observed weather and streamflow up to just before the forecast start time and assuming a

one-step-ahead weather forecast that turns out to perfectly match observations. In future work, we will extend ERRIS to multiple-step-ahead streamflow forecasting.

In this study we use the term "ensemble" to mean a set of equally probable realizations of future streamflow that represents the hydrological model uncertainty. The forecasts based on ERRIS are not typical probabilistic forecasts [*Gneiting and Katzfuss*, 2014], which explicitly provide the predictive distribution of future streamflow. For ERRIS, the probability distribution may be theoretically derived for one-step ahead forecasts based on the distributional assumption of model residuals. However, we can only obtain the predictive distribution of ERRIS forecasts at multiple step by means of Monte Carlo simulation.

The novelty of ERRIS is that it does not rely on a single complex error model, but runs a sequence of simple error models through multiple stages. We start with a very simple model of independent Gaussian residuals after data transformation to determine hydrological model parameters. At each subsequent stage, an error model is introduced to improve over the previous stage and to finalize the representation, including associated parameter values, of one particular statistical feature (bias, correlation in residuals or a non-Gaussian distribution). ERRIS progressively refines model features, focusing only on a small number of model parameters at each stage. This is achieved by estimating the values for a core set of parameters at each stage and holding them constant at subsequent stages. In doing so, ERRIS avoids the problems associated with parameter interactions that can occur under joint inference methods.

This paper is organized as follows. The ERRIS method is described in detail in Section 2. A case study is introduced in Section 3. Major results are presented in Section 4, followed by discussion and further results in Section 5. Conclusions are made in Section 6.

2. The error reduction and representation in stages (ERRIS) method

2.1. Model formulation

Stage 1: Transformation and hydrological modelling

We start from a simplified version of the seasonally invariant error model described by *Li et al.* [2013] to calibrate the hydrological model in the ERRIS method. At stage 1, we apply the log-sinh transformation [*Wang et al.*, 2012]

$$f(Q) = b^{-1}\log\{\sinh(a+bQ)\}, \quad (1)$$

where a and b are transformation parameters, to the raw values of streamflow Q. We assume at this stage that hydrological model forecast residuals are independent and, in the transformed space, follow a Gaussian distribution with a constant variance. The forecast variance in the original (untransformed) space is not a constant but is dependent on the magnitude of simulated streamflow through the back-transformation. The log-sinh transformation has been applied to a wide range of hydrological data [e.g. *Li et al.*, 2013; *Peng et al.*, 2014; *Robertson et al.*, 2013; *Shrestha et al.*, 2015; *Zhao et al.*, 2015] including extreme daily streamflow values [*Bennett et al.*, 2014b] to normalize data and stabilize variance, and has been shown to perform at least as well as other commonly used transformations [*Del Giudice et al.*, 2013; *Wang et al.*, 2012].

We denote the observed and simulated streamflows at day t by $Q(t)$ and $\tilde{Q}(t)$, respectively. The error model at Stage 1 is mathematically specified as

$$Z(t) = f(Q(t)) \tag{2}$$

$$\tilde{Z}_1(t) = f(\tilde{Q}(t)) \tag{3}$$

$$Z(t) \sim N\left(\tilde{Z}_1(t), \sigma_1^2\right) \tag{4}$$

where N denotes a Gaussian distribution of the model residuals in the transformed space at Stage 1, with mean $\tilde{Z}_1(t)$ and standard deviation σ_1. We will use similar notations (e.g. $\tilde{Q}$, Z, $\tilde{Z}$ and σ) for all stages in the ERRIS method, with stages distinguished by subscripts (i.e. 1, 2, 3, 4) . No autocorrelation within the forecast residuals is assumed at Stage 1. This avoids the potential parameter interference between the autocorrelation parameter and hydrological model parameters (e.g. parameters describing time persistence of the hydrograph) when the hydrological model is jointly calibrated with the error model. Stage 1 of the ERRIS method is summarized in Table 1. At the end of Stage 1, the simulated streamflow $\tilde{Q}(t)$ is taken as the forecast median of the ensemble streamflow forecast.

Stage 2: Linear bias correction

At Stage 1, we assume that the hydrological simulation is overall unbiased. However, the hydrological model often over-estimates low flows and under-estimates high flows. At Stage 2, we adopt a simple but effective bias-correction scheme firstly introduced by *Wang et al.* [2014]

to revise the the forecast value made at Stage 1. This bias correction describes the forecast bias in the transformed domain by a linear function. Because the bias-correction is applied to transformed data, it is able to cope with conditional biases (biases that vary with flow magnitude) that are often present in hydrological model simulations, even if these vary in a strongly non-linear way. We express the specific error model structure of Stage 2 as

$$\tilde{Z}_2(t) = c + d\tilde{Z}_1(t) \tag{5}$$

$$Z(t) \sim N\left(\tilde{Z}_2(t), \sigma_2^2\right) \tag{6}$$

where c and d represent the intercept and slope parameters of the bias correction and σ_2 denotes the standard deviation of the residuals at Stage 2. The slope parameter d allows much flexibility in the bias correction. When d equals 1, this bias correction becomes a simple additive correction. When d equals 0, the bias-correction forces the forecast to approach a constant (in additional to uncertainty). This may happen when the hydrological forecast performs worse than climatology (i.e. long-term average). When d is greater than 1, the bias-correction can correct the very strongly conditional biases, as might be found in ephemeral catchments.

At the end of Stage 2, the forecast median in the orginal space is revised to

$$\tilde{Q}_2(t) = f^{-1}\left(\tilde{Z}_2(t)\right), \tag{7}$$

where $f^{-1}(x) = b^{-1}\operatorname{arsinh}\{\exp(bx) - a\}$ is the back-transformation of the log-sinh transformation given in Equation (1).

Stage 3: AR updating

At Stage 3, we no longer assume that forecast residuals are independent, and use an AR-based error model to describe the correlation structure of forecast residuals. The AR-based error model enables the ERRIS method to correct forecast residuals based on the latest available observations of streamflow. Specifically, we assume that the forecast residuals at Stage 2 follow a restricted AR error model described by *Li et al.* [2015]. The error model at Stage 3 can be written as

$$\tilde{Z}_3(t) = \begin{cases} \tilde{Z}_2(t) + \rho\left(Z(t-1) - \tilde{Z}_2(t-1)\right) & \text{if } \left|\tilde{Q}_3^*(t) - \tilde{Q}_2(t)\right| \le \left|Q(t-1) - \tilde{Q}_2(t-1)\right| \\ f\left(\tilde{Q}_2(t) + Q(t-1) - \tilde{Q}_2(t-1)\right) & \text{otherwise} \end{cases} \quad (8)$$

$$Z(t) \sim N\left(\tilde{Z}_3(t), \sigma_3^2\right) \quad (9)$$

where $\tilde{Q}_3^*(t) = f^{-1}\left(\tilde{Z}_2(t) + \rho\left(Z(t-1) - \tilde{Z}_2(t-1)\right)\right)$ is the updated streamflow without applying the restriction, and ρ and σ_3 are the lag-1 autocorrelation parameter and the standard deviation of the residuals at Stage 3, respectively. *Li et al.* [2015] demonstrated that when AR models are applied to normalized residuals without restriction, over-correction of forecasts can occur, particularly at the peak or on the rise of a hydrograph. Equation (8) uses the restricted AR error model to reduce the tendency to over-correct forecasts. In Equation (8) the forecast median, denoted by $\tilde{Q}_3(t)$, is given by

$$\tilde{Q}_3(t) = \begin{cases} \tilde{Q}_3^*(t) & \text{if } \left|\tilde{Q}_3^*(t) - \tilde{Q}_2(t)\right| \le \left|Q(t-1) - \tilde{Q}_2(t-1)\right| \\ \tilde{Q}_2(t) + Q(t-1) - \tilde{Q}_2(t-1) & \text{otherwise} \end{cases}. \quad (10)$$

227 The forecast at Stage 3 updates $\tilde{Q}_2(t)$ based on the latest observed streamflow $Q(t-1)$ and its
228 difference from $\tilde{Q}_2(t-1)$. Therefore, more information (i.e. streamflow observations at the
229 previous time step) is required to generate streamflow forecasts at Stage 3 than at the previous
230 two stages.

231 *Stage 4: Residual distribution refinement*

232 In Section 4, we will demonstrate that the residuals after Stages 1 and 2 are well described
233 by Gaussian distributions, but the shape of the residual distribution after Stage 3
234 dramatically changes. In particular, the distribution of the residuals after Stage 3 looks more
235 peaked and has longer tails than a Gaussian distribution. The reason for the non-Gaussian
236 residuals after Stage 3 is as follows. The AR updating at Stage 3 is very effective in
237 correcting small residuals especially at hydrograph recession and therefore reducing
238 residuals to very small values. The updating, however, is not very effective around peaks,
239 where the residuals remain large even in the transformed space. This results in a centrally
240 peaked and long tailed distribution of residuals after Stage 3.

241 At Stage 4, we use a non-Gaussian distribution to describe the model residuals from Stage 3.
242 Several long-tailed distributions have been used in hydrological modelling studies, such as
243 the finite mixture distribution [*Schaefli et al.*, 2007; *Smith et al.*, 2010], the exponential
244 power distribution [*Schoups and Vrugt*, 2010] and Student's t-distribution [*Marshall et al.*,
245 2006]. In this study, we assume that the model residuals can be grouped into two categories
246 with respect to variance and thus choose a two-component Gaussian mixture distribution. It is
247 possible to use more than two components, but we will show in our case study that two

components are sufficient. We discuss the possibility of using other long-tailed distributions in Section 5.1.

Using a two-component Gaussian mixture distribution, we express the residual model at Stage 4 as

$$\tilde{Z}_4(t) = \tilde{Z}_3(t) \tag{11}$$

$$Z(t) \sim MN\left(\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right), \tag{12}$$

where $MN\left(\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right)$ represents a mixture of two Gaussian distributions $N\left(\tilde{Z}_4(t), \sigma_{4,1}^2\right)$ and $N\left(\tilde{Z}_4(t), \sigma_{4,2}^2\right)$ with weights w and $1-w$. The corresponding probability density function of $MN\left(\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right)$, denoted by $pdf\left(Z(t) \mid \tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right)$, can be explicitly written as a weighted sum of two Gaussian probability density functions

$$pdf\left(Z(t) \mid \tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right) = w\phi\left(Z(t) \mid \tilde{Z}_4(t), \sigma_{4,1}^2\right) + (1-w)\phi\left(Z(t) \mid \tilde{Z}_4(t), \sigma_{4,2}^2\right). \tag{13}$$

where ϕ is the probability density function (PDF) of a Gaussian distribution. We assume that $\sigma_{4,1} < \sigma_{4,2}$ to make the two components identifiable. This assumption implies that w represents the probability associated with the mixture component that has a smaller variance.

2.2. Model estimation

The maximum likelihood estimation [*Li et al.*, 2013; *Wang et al.*, 2009] is used to estimate model parameters at all four stages. Denote the parameter set as θ_S for Stage *S*. The likelihood functions for the four stages are given by

$$L_S(\theta_S) = \prod_t J_{z\to Q}\phi\left(Z(t) \mid \tilde{Z}_S(t), \sigma_S^2\right) \tag{14}$$

for $S = 1, 2, 3$, and

$$L_4(\theta_4) = \prod_t J_{z\to Q}\, pdf\left(Z(t) \mid \tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right) \tag{15}$$

where $J_{z\to Q} = 1/\tanh\{a + bQ(t)\}$ is the Jacobian determinant of the log-sinh transformation.

At Stage 1, the hydrological model parameters, transformation parameters (a and b) and the residual standard deviation (σ_1) are jointly estimated by maximizing the likelihood function. It is also possible to use a set of parameters already calibrated for the hydrological model (using a different objective, such as the least of the sum of squared errors) and estimate at Stage 1 only the transformation parameters and the residual standard deviation (see discussion in Section 5.2). At the end of Stage 1, the values of the hydrological parameters and the transformation parameters are concluded, without further changes in subsequent stages.

At Stage 2, the bias correction parameters (c and d) and the residual standard deviation (σ_2) are estimated by maximizing the likelihood function. At the end of Stage 2, the values of the bias

correction parameters are concluded. At Stage 3, the auto-correlation coefficient (ρ) and the residual standard deviation (σ_3) are estimated. At the end of Stage 3, the value of the auto-correlation coefficient is concluded. At Stage 4, the model residual parameters ($\sigma_{4,1}, \sigma_{4,2}$ and w) are finalized. The variances at Stages 1-3 (i.e. σ_1, σ_2 and σ_3) are not used to generate forecasts, but only for estimating parameters at corresponding stages. We use maximum likelihood at each stage to estimate parameters, and this requires us to specify the variance of residuals at each stage.

The Shuffled Complex Evolution (SCE) algorithm [*Duan et al.*, 1994] is used to maximize the log likelihood function at Stage 1, where a number of parameters are required to be calibrated. The Simplex algorithm [*Nelder and Mead*, 1965] is used in the likelihood-based calibration at other stages, where fewer parameters are present. We use different optimization algorithms because the Simplex algorithm is more computationally efficient when the number of parameters is small.

2.3. Model verification

We use several performance measures to evaluate the ensemble forecasts derived at each stage. The evaluation criteria suggested by *Engeland et al.* [2010] are used to test for important attributes of ensemble forecasts including *reliability*, *sharpness* and *efficiency*.

Reliability is often described as the property of statistical consistency, which allows ensemble forecasts to reproduce the frequency of an event. Reliability can be checked by the forecast probability integral transform (PIT) of streamflow observations, defined by

$$\pi_t = F_t(Q(t)) \tag{15}$$

where F_t is the forecast CDF of the streamflow at time t. In the case of zero flows, we use the pseudo PIT [*Wang and Robertson*, 2011], which is randomly generated from a uniform distribution with a range $[0, \pi_t]$. If a forecast is reliable, π_t follows a uniform distribution over [0,1]. We graphically examine π_t with the corresponding theoretical quantile of the uniform distribution using the PIT-uniform probability plot (or simply *PIT plot;* also called the predictive quantile quantile plot [*Renard et al.*, 2010]). The PIT plot is closely related to the rank histogram [*Gneiting et al.*, 2005; *Hamill*, 2001]. From our experience, the PIT plot is more suitable than the rank histogram for the experiments where observations are abundant (such as daily or sub-daily forecasting verification). A perfectly reliable forecast follows the 1:1 line. A Kolmogorov-Smirnov significant band can be included in the PIT plots to as a test of uniformity [*Laio and Tamea*, 2007]. In addition, PIT plots can be summarized by the α-index [*Renard et al.*, 2010], defined by

$$\alpha = 1 - \frac{2}{n}\sum_{t=1}^{n}\left|\pi_t^* - \frac{t}{n+1}\right|, \tag{16}$$

where π_t^* is the sorted π_t in increasing order. The α-index represents the total deviation of π_t^* from the corresponding uniform quantile (i.e., the tendency to deviate from the 1:1 line in PIT diagrams). The range of the α-index is from 0 (worst reliability) to 1 (perfect reliability).

Sharpness is a measure of the spread of the forecast probability distribution. Sharp forecasts with narrow forecast intervals are usually preferred by forecast users as they reduce the

range of possible outcomes that are anticipated – that is, it is easier to make decisions with sharp forecasts. However, if a sharp forecast is unreliable, it is overconfident and is likely to lead to poor decisions. Thus sharp forecasts are desirable, but only if the forecasts are also reliable. We use the average width of the 95% forecast intervals (AWCI) to indicate forecast sharpness [*Gneiting et al.*, 2007]. Wider forecast intervals suggest less sharp forecasts. In order to compare the sharpness across different catchments, we define a score relative AWCI with respect to a reference forecast

$$\text{Relative AWCI} = \frac{AWCI_{REF} - AWCI}{AWCI_{REF}}, \tag{17}$$

where $AWCI_{REF}$ is AWCI calculated from the reference forecast. The reference forecast in this study is generated by resampling historical streamflows. To issue a reference forecast for a given month/year (e.g. February 1999), we randomly draw a sample of 1000 daily streamflows that occur in that month (e.g. February) from other years (e.g. years other than 1999) with replacement. As only 14years of data are used in this study, the reference forecast for each month is more robust than the similar reference forecast for each day. The relative AWCI is unitless and the maximum is one, corresponding to the sharpest forecast.

The *Efficiency* (or accuracy) of a forecast is commonly used to assess deterministic (single-valued) forecasts. For the ensemble forecasts we generate here, we measure the efficiency with the well-known Nash-Sutcliffe efficiency (NSE) [*Nash and Sutcliffe*, 1970], calculated for the forecast mean. A greater value of NSE indicates a more accurate forecast mean and thus better forecast efficiency. We also use relative bias to assess how the forecast mean deviates from observations.

We evaluate the overall forecast skill with a skill score derived from the widely used continuous ranked probability score (CRPS) [*Gneiting and Katzfuss*, 2014; *Grimit et al.*, 2006; *Wang et al.*, 2009] (denoted by $CRPS\_SS$). CRPS is a negatively oriented score: a smaller value of CRPS indicates a better forecast. As with the relative AWCI, the skill score $CRPS\_SS$ is defined as the normalized version of CRPS with respect to a reference forecast

$$CRPS\_SS = \frac{CRPS_{REF} - CRPS}{CRPS_{REF}}, \quad (18)$$

where $CRPS_{REF}$ is CRPS calculated from the reference forecast (already defined for Equation (18), above). The maximum of $CRPS\_SS$ is 1, corresponding to a perfectly skillful forecast.

While a decomposition of CRPS is available that gives an indication of reliability [*Hersbach 2000*], we do not use this. PIT plots are a stronger test of reliability [*Candille and Talagrand*, 2005], and accordingly we focus on PIT plots to discuss reliability.

3. Case Study

3.1 Study region and data

We select six catchments in southeast Australia and three catchments in the United States (US) for this study (Figure 1), from a range of climatic and hydrological conditions. The streamflow data for the Australian catchments are obtained from the Catchment Water Yield Estimation Tool (CWYET) dataset [*Vaze et al.*, 2011]. The rainfall and potential evaporation data for the Australian catchments are taken from the Australian Water Availability Project (AWAP) dataset [*Jones et al.*, 2009]. All data for the US catchments are

358 taken from the Model Intercomparison Experiment (MOPEX) dataset [*Duan et al.*, 2006].
359 The Abercrombie and Emu catchments have many instances of zero flow (Table 2), and
360 accurate streamflow forecasting is particularly challenging for such dry catchments.
361 $AWCI_{REF}$ and $CRPS_{REF}$ for each catchment is given by Table 3.

362 **3.2 Cross-validation**

363 Daily streamflow is simulated with the GR4J rainfall-runoff model [*Perrin et al.*, 2003] and
364 then forecasted with ERRIS as described in Section 3. GR4J is a widely used conceptual
365 model that was designed to be as parsimonious as possible. Its four parameters describe the depth
366 of a production store (X1), groundwater exchange (X2), the depth of a routing store (X3) and the
367 length of unit hydrographs (X4). Forecasts are generated from "perfect" (observed)
368 deterministic rainfall forecasts at a lead time of one day (i.e., one time step ahead). All
369 results reported in this study are based on cross-validation unless specified. Cross-validation
370 allows us to generalize the forecast skill to data outside the sample period. Because of data
371 availability, we choose different study periods for Australian and US catchments. For
372 Australian catchments, data from 1990 to 1991 are used to warm up the hydrological model
373 and the data from 1992-2005 are used to generate a leave-two-years-out cross-validation
374 (i.e. effectively 14-fold cross-validation). For a particular year, we remove the streamflow
375 data from this year and the following year and apply ERRIS to forecast the streamflow for
376 the year. The removal of the data from the following year aims to minimize the impact of
377 streamflow memory on model performance. For US catchments, the data from 1979 to 1980
378 are used in the warm-up period and the data from 1981 to 1998 are used for a leave-two-
379 years-out cross-validation (i.e. effectively 18-fold cross-validation).

4. Results

Figure 2 compares forecasts at different stages for an example period. In this example, we generate daily streamflow forecasts for the Mitta Mitta catchment in the period between 01/07/2000 to 31/12/2000. The forecast mean and the 95% forecast interval are plotted against observations. The forecast at Stage 1 (the base hydrological model forecast) frequently over-estimates low flows, such as in the period between July and September. For high flow periods (e.g. October), the forecast mean is generally more accurate but virtually all observations lie within the 95% forecast intervals, suggesting that the forecast intervals are too wide (i.e., the forecasts may be underconfident). The forecast mean at Stage 2 is closer to the observations and the 95% forecast intervals tend to be narrower. Stage 2 tends to overestimate high flows less than Stage 1, but introduces the problem of underestimating high flows in some instances (e.g. September).

The AR error updating applied in Stage 3 significantly reduces the forecast residuals, as we expect given that streamflows are usually heavily autocorrelated. The forecasts at Stage 3 are not only more accurate but also more certain, indicated by the considerably narrower 95% forecast intervals. The differences between Stage 3 and Stage 4 are not evident in the time-series plots, in essence because Stage 4 is an attempt to address issues of reliability, which is difficult to see when forecast intervals are so narrow. We give a detailed view of changes to reliability at each stage below.

Figure 3 summarizes the performance at each stage for all catchments, and generally confirms the improvements in performance at each stage observed in Figure 2. In general, Stage 1 and Stage 2 are similarly efficient (Figure 3b), skillful (Figure 3c), sharp (Figure 3d) and reliable (Figure 3e). As we expect, Stage 2 forecasts are consistently less biased than Stage 1 (Figure 3a)

403 (except for the Hope catchment, where many instances of zero flow occur; see Table 2). Stage 3
404 is generally much more efficient and skillful than Stage 1 and Stage 2. A partial exception to this
405 is the Abercrombie catchment, which is less efficient at Stage 3 than Stage 2. The Abercrombie
406 catchment experiences low (to zero) flows, but is also punctuated by abrupt high flows. Stage 3
407 is based on the time persistence of the residuals and may introduce more errors when flows
408 change abruptly, which sometimes occurs in the Abercrombie catchment. In addition, residuals
409 tend to be larger at higher flows and because NSE is a measure of squared residuals, it tends to
410 give more weights to residuals at high flows. This causes the Abercrombie Stage 3 forecasts to
411 be less efficient than those of Stage 2.
412 As we expect, Stage 3 forecasts are notably sharper than those at Stage 2 (Figure 3d). However,
413 this sharpness is not supported by reliability: Stage 3 forecasts tend to be much less reliable than
414 all other stages (Figure 3e). Figure 4 illustrates the reliability of the forecasts at each stage in
415 more detail with the PIT plots. As PIT values are highly autocorrelated, we have to "thin" them
416 in order to make the Kolmogorov-Smirnov significant band applicable [*Zhao et al.*, 2015]. We
417 generate PIT plots from every 30-th forecast to eliminate the autocorrelation. The PIT plots show
418 that the forecasts at the first two stages are reliable (as with the α-index in Figure 3e). However,
419 for Stage 3 the points on the PIT plots deviate substantially from the 1:1 line, with a clear S-
420 shape pattern for almost all catchments (the exception is the Tarwin catchment). A traditional
421 interpretation of this S-shape is that the forecasts are underconfident [*Laio and Tamea*, 2007].
422 However, in this case, the S-shape is caused by the high level of kurtosis in the distribution of the
423 residuals, as we will show below. The α-index from Stage 3 is smaller than those from stages 1
424 and 2 (the Tarwin catchment is the only exception), confirming the lack of the reliability at Stage
425 3. Stage 4 consistently improves the reliability of the forecast after the AR updating. The PIT

426 plot at Stage 4 is much closer to the 1:1 line than that at Stage 3 and this is reflected by the α-
427 index, which increases for all catchments. Stage 4 corrects the underconfident forecasts from
428 Stage 3 and slightly decreases the sharpness from Stage 3 (Figure 3d).
429 At Stage 3, unreliable forecasts are caused by representing the model residual by an
430 inappropriate (Gaussian) probability distribution. We compare the underlying density of the
431 model residuals at Stage 3, $\varepsilon(t) = Z_3(t) - \tilde{Z}_3(t)$ (fitted by the nonparametric density estimation),
432 with the fitted parametric densities for different distributions in Figure 5. The fitted Gaussian
433 density is flatter than the underlying density of $\varepsilon(t)$ in order to match the tails for each
434 catchment. This suggests that the residual distribution is more peaked and has longer tails than
435 the Gaussian distribution. As we have seen above, forecast residuals are, in general, dramatically
436 reduced by the AR error updating. Unfortunately, this reduction in residuals does not occur at all
437 events, especially where abrupt changes in flow occur (and hence the assumption of strong
438 autocorrelation breaks down). Thus the magnitude of the forecast residuals at Stage 3 for a small
439 proportion of events is large relative to the majority of events. As we have seen, the practical
440 implication of the dichotomous behavior of the residuals is that their distribution is still bell-
441 shaped and symmetric but is peakier and has a much longer tail than the Gaussian distribution.
442 The Gaussian mixture distribution treats model residuals as two groups with different variances.
443 The Gaussian mixture distribution is able to capture the peak and tails of the underlying residual
444 density for all catchments, resulting in reliable ensemble forecasts that also have a highly
445 accurate forecast mean. As we note in the introduction, however, other distributions have also
446 been used to describe “peaky” data, and we explore these in the next section.

447 To provide a basis for any future comparisons with this study, we include example parameter
448 values for each stage in Table 4 (derived by calibrating each stage to the full set of data – i.e.
449 without cross-validation). We note that: 1) the variance parameter at Stage 3 is always much
450 smaller than at Stage 1 and Stage 2, which leads to the dramatic reduction in the width of
451 forecast intervals at this stage; and 2) that the W parameter that weights the component of the
452 Gaussian mixture distribution with smaller variance is always greater than 0.5, confirming that
453 the majority of residuals take a narrow range of values as we have described.

454 **5. Further results**

455 **5.1 Testing an alternative residual distribution**

456 It is possible to use long-tailed distributions other than the Gaussian mixture distribution at Stage
457 4. For example, Student's t-distribution is a simple long-tailed distribution that has been used in
458 hydrological modelling [e.g. *Marshall et al.*, 2006]. In this section we investigate whether
459 Student's t-distribution is a viable alternative to the Gaussian mixture distribution at Stage 4. To
460 do this, we modify the model residual in Equation (12) as follows

461 $$Z(t) = \tilde{Z}_4(t) + r\xi(t), \qquad (19)$$

462 Where $\xi(t)$ is assumed to independently follow a Student's t-distribution with ν degrees of
463 freedom, and r is a scale parameter describing the spread and variation of the model residuals.

464 We first examine how well Student's t-distribution can fit the residual distribution at Stage 4 for
465 all nine catchments (Figure 5). High peaks and long tails of the residual densities can be captured
466 reasonably well by Student's t-distribution for nearly all catchments. The fitted densities of

467 Student's t-distribution appear more "peaked" for most catchments than those of the Gaussian
468 mixture distribution, which is originally used at Stage 4. Figure 6 further investigates how
469 Student's t-distribution can fit the upper quantile of the model residuals. There is a clear
470 tendency of Student's t-distribution to overestimate the upper quantile (e.g. 98% or higher) of the
471 model residuals (especially for the Australian catchments). These upper quantiles are more
472 accurately estimated by the Gaussian mixture distribution. This implies that Student's t-
473 distribution often has tails that are too long. We note, however, that if the ERRIS method is
474 tested on other catchments, it is possible that Student's t-distribution may describe the residuals
475 better than the Gaussian mixture distribution in some cases.

476 A disadvantage of the very long tail of Student's t distribution is that it can be problematic for
477 operational forecasting. The degrees of freedom, ν, determines how heavy the tails of Student's
478 t-distribution are. Table 5 presents the two calibrated parameters (i.e. ν and r) for all
479 catchments. Calibrated $_{\nu}$ values are less than 2 for eight out of nine catchments. The exception is
480 the Hope catchment, and even here the calibrated ν is very close to 2. It is well know that for
481 degrees of freedom less than 2, Student's t-distribution is so heavy-tailed that the variance is
482 infinite (if $1<\nu\leq 2$) or even undefined (if $\nu\leq 1$). This is obviously undesirable for operational
483 forecasting: it can cause a few forecast ensemble members to be so large that the forecast mean
484 becomes implausibly large. Figure 7 compares the forecast mean with observations if the model
485 residual is revised as Equation (19). In all catchments, in some cases forecast mean values are
486 unrealistically large even as observations are relatively small. Student's t-distribution is thus
487 prone to be too long-tailed to be practically implemented. Therefore, we do not recommend

using Student's t-distribution to describe the residual distribution at Stage 4, and advocate the Gaussian mixture distribution as a practical alternative.

5.2 Testing an alternative calibration of the hydrological model

In this study, we apply a likelihood-based calibration at Stage 1 to derive the distribution of the forecast residuals. However, in operational practice forecasters may prefer to use their own methods for calibrating hydrological models (or it may be onerous to recalibrate large numbers of hydrological models, whatever method is used). It is possible to simply 'bolt on' the ERRIS method to existing hydrological models. We simply need to calibrate the transformation parameters and the model residual standard deviation at Stage 1 while fixing the hydrological parameters to those already calibrated. We demonstrate this by first calibrating hydrological models with a simple least-squares objective. We then apply the ERRIS method and repeat the cross-validation analysis.

Figure 8, an analog to Figure 3, summarizes forecast performance when the hydrological model is calibrated to a least-squares objective. The least-squares calibration essentially maximizes NSE as an objective, but the corresponding cross-validated NSE is not necessarily always greater than that of the likelihood-based calibration. The forecast performance from the two different calibrations can differ markedly at Stage 1, but is largely similar after the AR error updating at Stage 3 and Stage 4. Thus ERRIS is flexible enough to accommodate existing hydrological models.

Figure 9, an analog to Figure 4, compares the PIT plots for different catchments when the hydrological model is least-squares calibrated. The main change is that the forecasts at Stage 1

509 are no longer reliable in many instances. This is caused by the least-squares calibration, which
510 does not ensure the forecast residuals are Gaussian (even after the log-sinh transformation). The
511 PIT plots derived from Stage 2 and Stage 3 in Figure 9 show a very similar pattern to their
512 counterparts in Figure 4. It suggests that poor reliability at Stage 3 occurs irrespective of the
513 calibration strategy employed for the hydrological model. As with Figure 4, Figure 9 shows the
514 Gaussian mixture distribution used at Stage 4 effectively ameliorates the problems with the
515 reliability of Stage 3.

516 **6. Discussion**

517 There are several advantages of using a multi-stage error model compared to a single complex
518 error model. (1) The parameter estimation in ERRIS is relatively simple, and hence
519 computationally efficient. Only a small number of parameters are estimated at each stage. Joint
520 parameter estimations associated with a single complicated error model are often more
521 computationally demanding. (2) Interference between parameters is minimized. The parameters
522 of a single complex model can confound each other and the contribution of one parameter can
523 sometimes be explained by others. For example, the hydrological model parameters describing
524 soil moisture storage capacity may interfere strongly with the error parameters describing bias.
525 Interference between parameters can make the parameter estimation unstable, because more than
526 one set of parameters can achieve a similar objective function value, and thus over-fit
527 parameters. (3) In operational forecasting it is often important that individual components of the
528 forecasting model can function independently. For example, if forecasts are issued to long lead
529 times, the influence of an AR model diminishes as lead time extends. Thus forecasts at long lead
530 times rely strongly on the hydrological model (and, in our case, a bias-correction) to be plausible,

even with perfect meteorological forcings. If all parameters are estimated jointly, it is difficult to guarantee that each component of a forecasting model can operate independently. In addition, because stages are independent, it is possible to change a stage without affecting other stages, making the ERRIS approach easy to extend or modify.

This paper is aimed at developing a staged error model suitable for eventual use in an operational ensemble forecasting system. We have focused on presenting the theoretical underpinnings of this approach, and have limited its testing to forecasting with 'perfect' (observed) rainfall forecasts at a lead time of one day. Operational systems routinely forecast to long lead times, and use uncertain rainfall forecasts to force hydrological models. In future work we will extend the validation of this model to forecast multiple lead times, and couple the ERRIS approach with reliable ensemble rainfall forecasts [*Robertson et al.*, 2013; *Shrestha et al.*, 2015].

The staged approach of ERRIS sets it apart from several predecessors, for example the hydrological uncertainty processor (HUP) and the dynamic uncertainty model by regression on absolute error (DUMBRAE). HUP is a Bayesian forecasting system to produce probabilistic streamflow forecasts [*Kelly and Krzysztofowicz*, 1997; *Krzysztofowicz*, 1999; 2001; *Krzysztofowicz and Kelly*, 2000; *Reggiani et al.*, 2009; *Todini*, 2008]. HUP and ERRIS have some similarities: (1) both are post-processors of deterministic hydrological models for hydrological uncertainty quantification; (2) both apply transformation to normalize data; (3) both use a linear regression in the transformed space for bias correction; (4) both use an autoregressive model to update hydrological simulation. However, ERRIS differs fundamentally from HUP by being implemented in stages. As we have noted, the staged approach avoids unwanted the interaction between parameters, and ensure the base

hydrological model performs as strongly as possible. In addition, some other technical advances distinguish ERRIS from HUP. For instance, ERRIS applies a restricted autoregressive model in order to avoid the possible overcorrection from the ordinary autoregressive model used in HUP. ERRIS uses a mixture of two Gaussian distributions for the residual distribution, which is more flexible than a Gaussian distribution used in HUP to describe the peak, shoulder and tail of the distribution.

Pianosi and Raso [2012] developed DUMBRAE to quantify predictive uncertainty of deterministic hydrological models. Unlike ERRIS, DUMBRAE does not apply data transformation and considers an error model in the original space. To deal with heteroscedastic residual errors, DUMBRAE explicitly formulates the error variance as a function of time series of absolute hydrological model errors and several independent predictors (such as precipitation). The dynamic variance model of DUMBRAE is an interesting alternative to the method we have presented here. As with HUP, another major difference between ERRIS and DUMBARE is staged error modelling that allows ERRIS to characterize the forecast error in stages and to avoid potential parameter interference and ensure robust performance of the base hydrological model.

7. Summary and conclusions

In this study, we introduce the error reduction and representation in stages (ERRIS) method to update errors and quantify uncertainty in streamflow forecasts. The first stage of ERRIS employs a simple error model that assumes independent Gaussian residuals after the log-sinh transformation. The second stage applies a bias-correction that is able to correct conditional and unconditional biases, including the sometimes strongly non-linear biases that occur in ephemeral catchments. The third stage exploits autocorrelation in residuals with an AR model to

dramatically reduce forecast residuals, but this results in unreliable ensemble forecasts. In the fourth stage a Gaussian mixture distribution is used to describe the residuals, resulting in ensemble forecasts that are both highly accurate and very reliable. Based on extensive validation of ERRIS, the accuracy of the forecast mean is slightly improved by the bias correction at Stage 2 and is considerably improved by the updating at Stage 3. The reliability of the forecasts at Stage 3 becomes a problem, because the shape of the residual distribution dramatically changes. The revision of the residual distribution at Stage 4 is effective for representing non-Gaussian residuals and leading to highly reliable forecasts. The Gaussian mixture distribution is showed to be more suitable than the Student's t distribution for describing the residuals after updating. ERRIS was designed with operational forecasting in mind, and we have shown that it is flexible enough to adapt to existing calibrated hydrological models.

Acknowledgements

This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Bureau of Meteorology and CSIRO Land & Water Flagship. We would like to thank Andrew Schepen for valuable suggestions to improve quality of the manuscript.

590 **REFERENCES**

591 Ajami, N. K., Q. Y. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian
592 multimodel combination framework: Confronting input, parameter, and model structural
593 uncertainty in hydrologic prediction, *Water Resour Res*, *43*(1). doi: 10.1029/2005wr004745
594 Alfieri, L., P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, and F. Pappenberger
595 (2013), GloFAS - global ensemble streamflow forecasting and flood early warning, *Hydrol Earth*
596 *Syst Sc*, *17*(3), 1161-1175. doi: 10.5194/hess-17-1161-2013
597 Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter
598 estimation and inference in conceptual rainfall-runoff modeling, *Water Resour Res*, *37*(4), 937-
599 947. doi: 10.1029/2000wr900363
600 Bennett, J. C., D. E. Robertson, D. L. Shrestha, Q. J. Wang, D. Enever, P. Hapuarachchi, and N.
601 K. Tuteja (2014a), A System for Continuous Hydrological Ensemble to lead times of 9 days
602 Forecasting (SCHEF), *J Hydrol*, *519*, 2832-2846. doi:
603 Bennett, J. C., D. E. Robertson, D. L. Shrestha, Q. J. Wang, D. Enever, P. Hapuarachchi, and N.
604 K. Tuteja (2014b), The challenge of forecasting high streamflows 1-3 months in advance with
605 lagged climate indices in southeast Australia, *Nat Hazard Earth Sys*, *14*(2), 219-233. doi:
606 10.5194/nhess-14-219-2014
607 Candille, G., and O. Talagrand (2005), Evaluation of probabilistic prediction systems for a scalar
608 variable, *Q J Roy Meteor Soc*, *131*(609), 2131-2150. doi: 10.1256/qj.04.71
609 Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann (2013),
610 Improving uncertainty estimation in urban hydrological modeling by statistically describing bias,
611 *Hydrol. Earth Syst. Sci.* , *17*, 4209-4225. doi: 10.5194/hess-17-4209-2013

612 Demargne, J., et al. (2014), The Science of NOAA's Operational Hydrologic Ensemble Forecast
613 Service, *B Am Meteorol Soc*, *95*(1), 79-98. doi: 10.1175/bams-d-12-00081.1
614 Diskin, M. H., and E. Simon (1977), A procedure for the selection of objective functions for
615 hydrologic simulation models, *J Hydrol*, *34*(1–2), 129-149. doi: 10.1016/0022-1694(77)90066-X
616 Duan, Q. Y., S. Sorooshian, and V. K. Gupta (1994), Optimal Use of the Sce-Ua Global
617 Optimization Method for Calibrating Watershed Models, *J Hydrol*, *158*(3-4), 265-284. doi:
618 10.1016/0022-1694(94)90057-4
619 Duan, Q. Y., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of
620 science strategy and major results from the second and third workshops, *J Hydrol*, *320*(1-2), 3-
621 17. doi: 10.1016/j.jhydrol.2005.07.031
622 Engeland, K., B. Renard, I. Steinsland, and S. Kolberg (2010), Evaluation of statistical models
623 for forecast errors from the HBV model, *J Hydrol*, *384*(1-2), 142-155. doi:
624 10.1016/j.jhydrol.2010.01.018
625 Evin, G., D. Kavetski, M. Thyer, and G. Kuczera (2013), Pitfalls and improvements in the joint
626 inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water*
627 *Resour Res*, *49*(7), 4518-4524. doi: 10.1002/wrcr.20284
628 Evin, G., M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera (2014), Comparison of joint
629 versus postprocessor approaches for hydrological uncertainty estimation accounting for error
630 autocorrelation and heteroscedasticity, *Water Resour Res*, *50*(3), 2350-2375. doi:
631 10.1002/2013WR014185
632 Gneiting, T., and M. Katzfuss (2014), Probabilistic Forecasting, *Annu Rev Stat Appl*, *1*, 125-151.
633 doi: 10.1146/annurev-statistics-062713-085831

634 Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and
635 sharpness, *J Roy Stat Soc B*, *69*, 243-268. doi: 10.1111/j.1467-9868.2007.00587.x
636 Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005), Calibrated probabilistic
637 forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly*
638 *Weather Review*, *133*(5), 1098-1118. doi: Doi 10.1175/Mwr2904.1
639 Gragne, A. S., A. Sharma, R. Mehrotra, and K. Alfredsen (2015), Improving real-time inflow
640 forecasting into hydropower reservoirs through a complementary modelling framework, *Hydrol.*
641 *Earth Syst. Sci.*, *19*(8), 3695-3714. doi: 10.5194/hess-19-3695-2015
642 Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson (2006), The continuous ranked
643 probability score for circular variables and its application to mesoscale forecast ensemble
644 verification, *Q J Roy Meteor Soc*, *132*(621), 2925-2942. doi: 10.1256/qj.05.235
645 Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts,
646 *Monthly Weather Review*, *129*(3), 550-560. doi: 10.1175/1520-
647 0493(2001)129<0550:IORHFV>2.0.CO;2
648 Jones, D. A., W. Wang, and R. Fawcett (2009), High-quality spatial climate data-sets for
649 Australia, *Australian Meteorological and Oceanographic Journal*, *58*, 233-248. doi:
650 Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in
651 hydrological modeling: 1. Theory, *Water Resour Res*, *42*(3). doi: 10.1029/2005wr004368
652 Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in
653 hydrological modeling: 2. Application, *Water Resour Res*, *42*(3). doi: 10.1029/2005wr004376
654 Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in
655 hydrology, *Stoch Hydrol Hydraul*, *11*(1), 17-31. doi: Doi 10.1007/Bf02428423

656 Krzysztofowicz, R. (1997), Transformation and normalization of variates with specified
657 distributions, *J Hydrol*, *197*(1-4), 286-292. doi: 10.1016/S0022-1694(96)03276-3
658 Krzysztofowicz, R. (1999), Bayesian theory of probabilistic forecasting via deterministic
659 hydrologic model, *Water Resour Res*, *35*(9), 2739-2750. doi: Doi 10.1029/1999wr900099
660 Krzysztofowicz, R. (2001), The case for probabilistic forecasting in hydrology, *J Hydrol*, *249*(1-
661 4), 2-9. doi: Doi 10.1016/S0022-1694(01)00420-6
662 Krzysztofowicz, R., and K. S. Kelly (2000), Hydrologic uncertainty processor for probabilistic
663 river stage forecasting, *Water Resour Res*, *36*(11), 3265-3277. doi: 10.1029/2000WR900108
664 Kuczera, G. (1983), Improved Parameter Inference in Catchment Models .1. Evaluating
665 Parameter Uncertainty, *Water Resour Res*, *19*(5), 1151-1162. doi: 10.1029/WR019i005p01151
666 Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error
667 analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent
668 parameters, *J Hydrol*, *331*(1-2), 161-177. doi: 10.1016/j.jhydrol.2006.05.010
669 Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous
670 hydrological variables, *Hydrol Earth Syst Sc*, *11*(4), 1267-1277. doi: 10.5194/hess-11-1267-2007
671 Li, M., Q. J. Wang, and J. C. Bennett (2013), Accounting for seasonal dependence in
672 hydrological model errors and prediction uncertainty, *Water Resour Res*, *49*(9), 5913-5929. doi:
673 10.1002/wrcr.20445
674 Li, M., Q. J. Wang, J. C. Bennett, and D. E. Robertson (2015), A strategy to overcome adverse
675 effects of autoregressive updating of streamflow forecasts, *Hydrol. Earth Syst. Sci.*, *19*(1), 1-15.
676 doi: 10.5194/hess-19-1-2015
677 Marshall, L., A. Sharma, and D. Nott (2006), Modeling the catchment via mixtures: Issues of
678 model specification and validation, *Water Resour Res*, *42*(11). doi: 10.1029/2005WR004613

679 Montanari, A., and A. Brath (2004), A stochastic approach for assessing the uncertainty of
680 rainfall-runoff simulations, *Water Resour Res*, *40*(1). doi: Artn W01106
681 10.1029/2003wr002540
682 Moradkhani, H., K. L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of
683 hydrologic model states and parameters: Sequential data assimilation using the particle filter,
684 *Water Resour Res*, *41*(5). doi: 10.1029/2004wr003604
685 Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I
686 — A discussion of principles, *J Hydrol*, *10*(3), 282-290. doi: 10.1016/0022-1694(70)90255-6
687 Nelder, J. A., and R. Mead (1965), A Simplex Method for Function Minimization, *The Computer*
688 *Journal*, *7*(4), 308-313. doi: 10.1093/comjnl/7.4.308
689 Peng, Z. L., Q. J. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. R. Wang
690 (2014), Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal
691 precipitation over China, *J Geophys Res-Atmos*, *119*(12), 7116-7135. doi:
692 10.1002/2013JD021162
693 Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for
694 streamflow simulation, *J Hydrol*, *279*(1-4), 275-289. doi: 10.1016/S0022-1694(03)00225-7
695 Pianosi, F., and L. Raso (2012), Dynamic modeling of predictive uncertainty by regression on
696 absolute errors, *Water Resour Res*, *48*. doi: Artn W03516
697 10.1029/2011wr010603
698 Reggiani, P., M. Renner, A. H. Weerts, and P. H. A. J. M. van Gelder (2009), Uncertainty
699 assessment via Bayesian revision of ensemble streamflow predictions in the operational river
700 Rhine forecasting system, *Water Resour Res*, *45*. doi: Artn W02428
701 10.1029/2007wr006758

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 Resour Res*, *46*. doi: 10.1029/2009wr008328

Robertson, D. E., D. L. Shrestha, and Q. J. Wang (2013), Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, *Hydrol Earth Syst Sc*, *17*(9), 3587-3603. doi: 10.5194/hess-17-3587-2013

Schaefli, B., D. B. Talamba, and A. Musy (2007), Quantifying hydrological modeling errors through a mixture of normal distributions, *J Hydrol*, *332*(3-4), 303-315. doi: 10.1016/j.jhydrol.2006.07.005

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 Resour Res*, *46*. doi: W10531, 10.1029/2009wr008933

Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. J. Wang (2015), Improving Precipitation Forecasts by Generating Ensembles through Postprocessing, *Monthly Weather Review*. doi: 10.1175/MWR-D-14-00329.1

Shrestha, D. L., D. E. Robertson, Q. J. Wang, T. C. Pagano, and H. A. P. Hapuarachchi (2013), Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose, *Hydrol Earth Syst Sc*, *17*(5), 1913-1931. doi: 10.5194/hess-17-1913-2013

Smith, T., A. Sharma, L. Marshall, R. Mehrotra, and S. Sisson (2010), Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resour Res*, *46*. doi: 10.1029/2010wr009514

724 Solomatine, D. P., and D. L. Shrestha (2009), A novel method to estimate model uncertainty
725 using machine learning techniques, *Water Resour Res*, *45*. doi: Artn W00b11
726 10.1029/2008wr006839
727 Sorooshian, S., and J. A. Dracup (1980), Stochastic Parameter-Estimation Procedures for
728 Hydrologic Rainfall-Runoff Models - Correlated and Heteroscedastic Error Cases, *Water Resour*
729 *Res*, *16*(2), 430-442. doi: 10.1029/WR016i002p00430
730 Thielen, J., J. Bartholmes, M. H. Ramos, and A. de Roo (2009), The European Flood Alert
731 System - Part 1: Concept and development, *Hydrol Earth Syst Sc*, *13*(2), 125-140. doi:
732 10.5194/hess-13-125-2009
733 Thyer, M., G. Kuczera, and Q. J. Wang (2002), Quantifying parameter uncertainty in stochastic
734 models using the Box-Cox transformation, *Journal of Hydrology*, *265*(1-4), 246-257. doi:
735 10.1016/S0022-1694(02)00113-0
736 Todini, E. (2008), A model conditional processor to assess predictive uncertainty in flood
737 forecasting, *International Journal of River Basin Management*, *6*(2), 123-137. doi:
738 10.1080/15715124.2008.9635342
739 Vaze, J., J. M. Perraud, J. Teng, F. H. S. Chiew, B. Wang, and Z. Yang (2011), Catchment Water
740 Yield Estimation Tools (CWYET), in *the 34th World Congress of the International Association*
741 *for Hydro- Environment Research and Engineering: 33rd Hydrology and Water Resources*
742 *Symposium and 10th Conference on Hydraulics in Water Engineering*, edited by E. Valentine, C.
743 Apelt, J. Ball, H. Chanson and J. Sargison, pp. 1554-1561, Engineers Australia, Brisbane. doi:
744 Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved
745 treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization
746 and data assimilation, *Water Resour Res*, *41*(1). doi: 10.1029/2004wr003059

747 Wang, Q. J., and D. E. Robertson (2011), Multisite probabilistic forecasting of seasonal flows for
748 streams with zero value occurrences, *Water Resour Res*, *47*. doi: W02546,
749 10.1029/2010WR009333
750 Wang, Q. J., D. E. Robertson, and F. H. S. Chiew (2009), A Bayesian joint probability modeling
751 approach for seasonal forecasting of streamflows at multiple sites, *Water Resour Res*, *45*. doi:
752 10.1029/2008WR007355
753 Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel (2012), A log-sinh transformation
754 for data normalization and variance stabilization, *Water Resour Res*, *48*. doi: W05514,
755 10.1029/2011WR010973
756 Wang, Q. J., J. C. Bennett, A. Schepen, D. E. Robertson, Y. Song, and M. Li (2014), FoGSS - A
757 model for generating forecast guided stochastic scenarios of monthly streamflows out to 12
758 months. *Rep.*, CSIRO Water for a Healthy Country Flagship, Highett, Australia.
759 Xiong, L. H., and K. M. O'Connor (2002), Comparison of four updating models for real-time
760 river flow forecasting, *Hydrolog Sci J*, *47*(4), 621-639. doi: 10.1080/02626660209492964
761 Yang, J., P. Reichert, K. C. Abbaspour, and H. Yang (2007), Hydrological modelling of the
762 chaohe basin in china: Statistical model formulation and Bayesian inference, *J Hydrol*, *340*(3-4),
763 167-182. doi: 10.1016/j.jhydrol.2007.04.006
764 Zhao, T., Q. J. Wang, J. C. Bennett, D. E. Robertson, Q. Shao, and J. Zhao (2015), Quantifying
765 predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model, *J*
766 *Hydrol*, *528*, 329-340. doi: 10.1016/j.jhydrol.2015.06.043

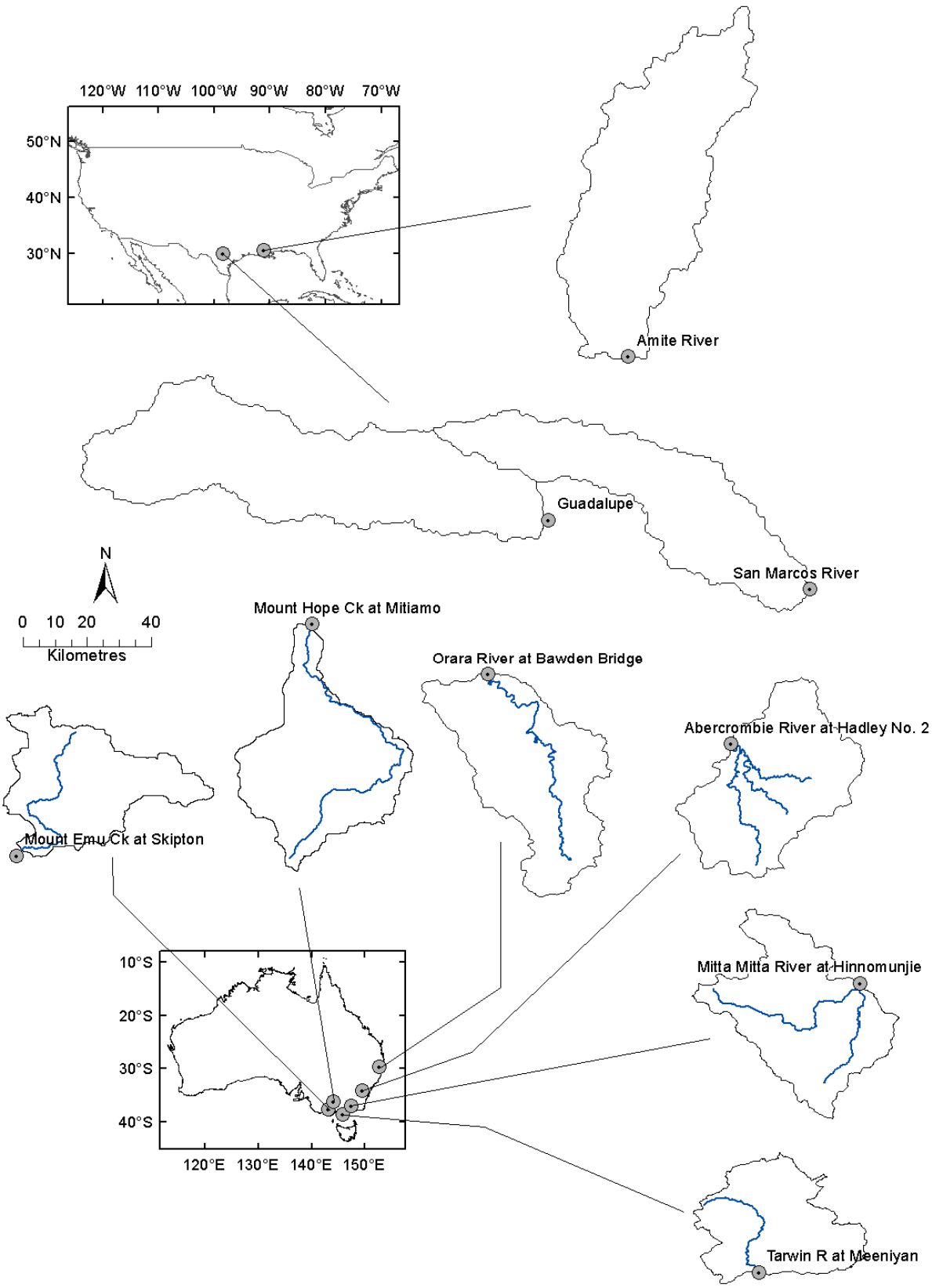
767 **Table of Figures**

769 Figure 2: An example of streamflow time-series plots for the Mitta Mitta catchment in the period
770 between 01/07/2000 and 31/12/2000.
771 Figure 3: Comparison of performance metrics for each catchment and each stage
772 Figure 4: Comparison of the cumulative probability distribution of the PIT at different stages
773 (light blue shaded strips indicate the 95% significant band of the Kolmogorov-Smirnov test.)
774 Figure 5: Comparison of the different probability density functions fitted to the model residuals
775 at Stage 3 for each catchment.
776 Figure 6: Comparison of the upper quantile of the model residuals fitted by different distributions
777 for each catchment.
778 Figure 7: Comparison of streamflow observations with streamflow forecast mean for each
779 catchment when the residual distribution is fitted by Student's t-distribution.
780 Figure 8: Same as Figure 3 but the hydrological model is calibrated by the least-squares method.
781 Figure 9: Same as Figure 4 but the hydrological model is calibrated by the least-squares method.
782
783

784

785 **Table of Tables**

792


120°W 110°W 100°W 90°W 80°W 70°W
50°N
40°N
30°N
Amite River
Guadalupe
San Marcos River
N
0 10 20 40
Kilometres
Mount Hope Ck at Mitiamo
Orara River at Bawden Bridge
Abercrombie River at Hadley No. 2
Mount Emu Ck at Skipton
Mitta Mitta River at Hinnomunjie
10°S
20°S
30°S
40°S
120°E 130°E 140°E 150°E
Tarwin R at Meeniyan

794 **Figure 1: Map of the catchments used in this study**
795

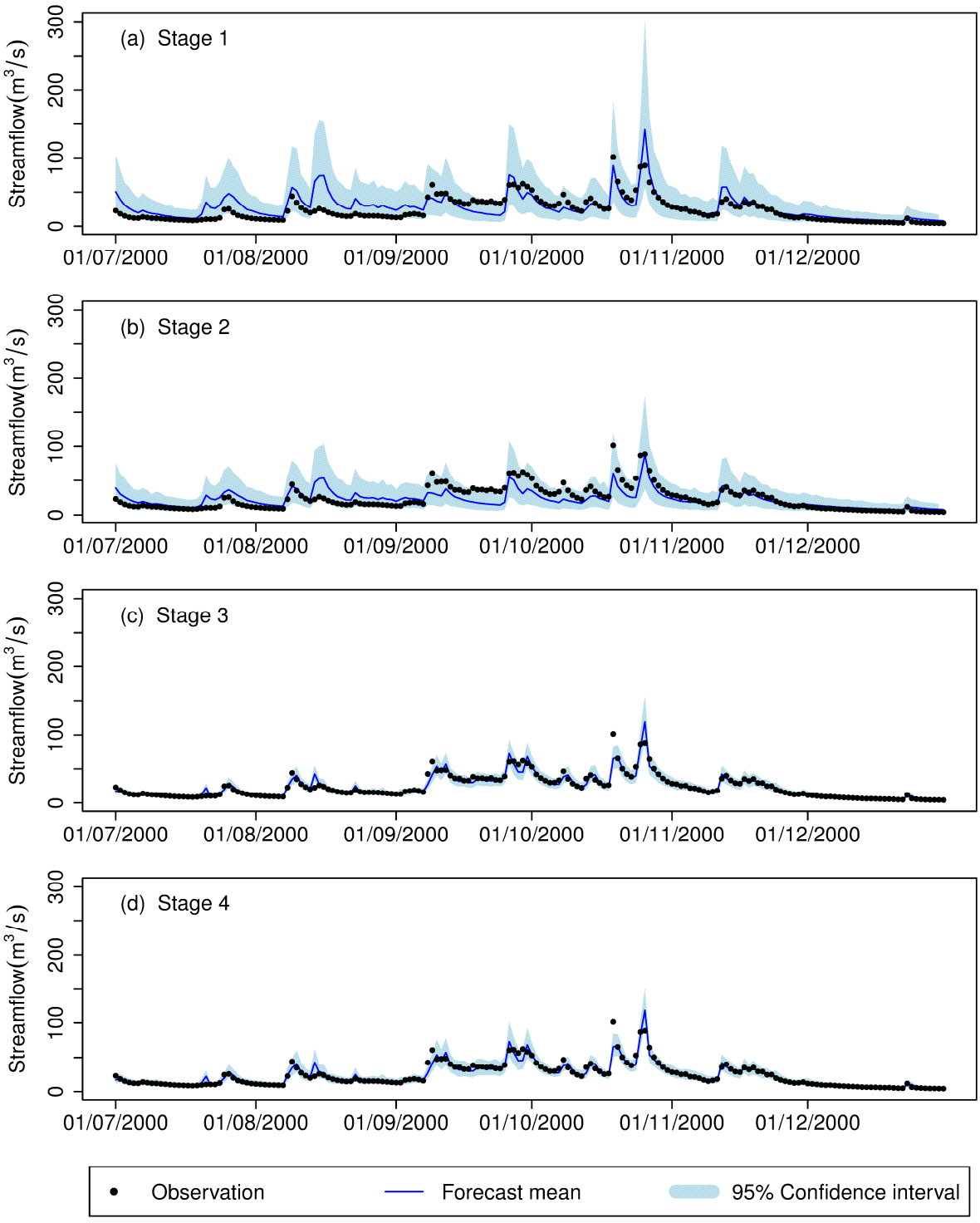
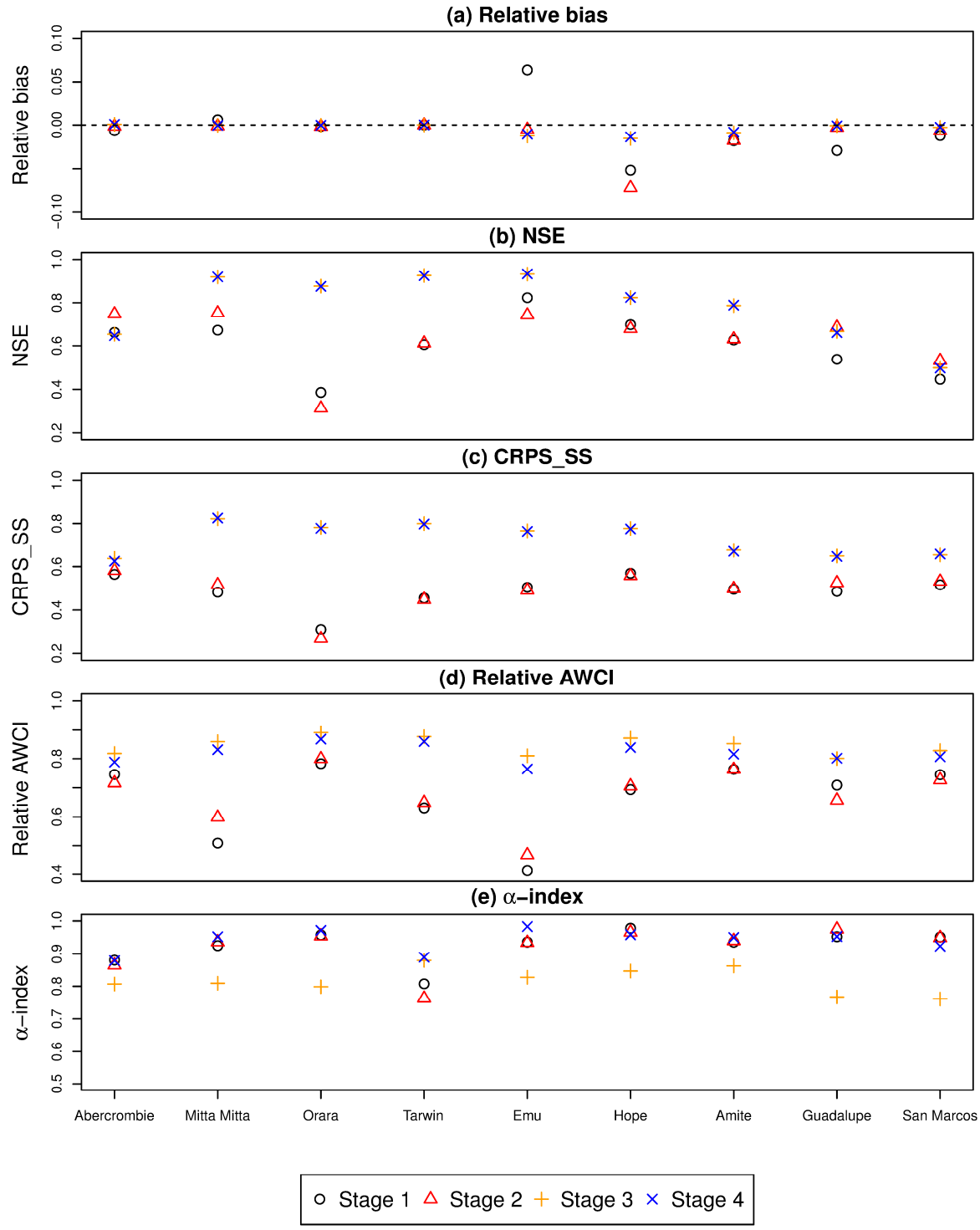


Figure 2: An example of streamflow time-series plots for the Mitta Mitta catchment in the period between 01/07/2000 and 31/12/2000.



801

802 **Figure 3: Comparison of performance metrics for each catchment and each stage**
803

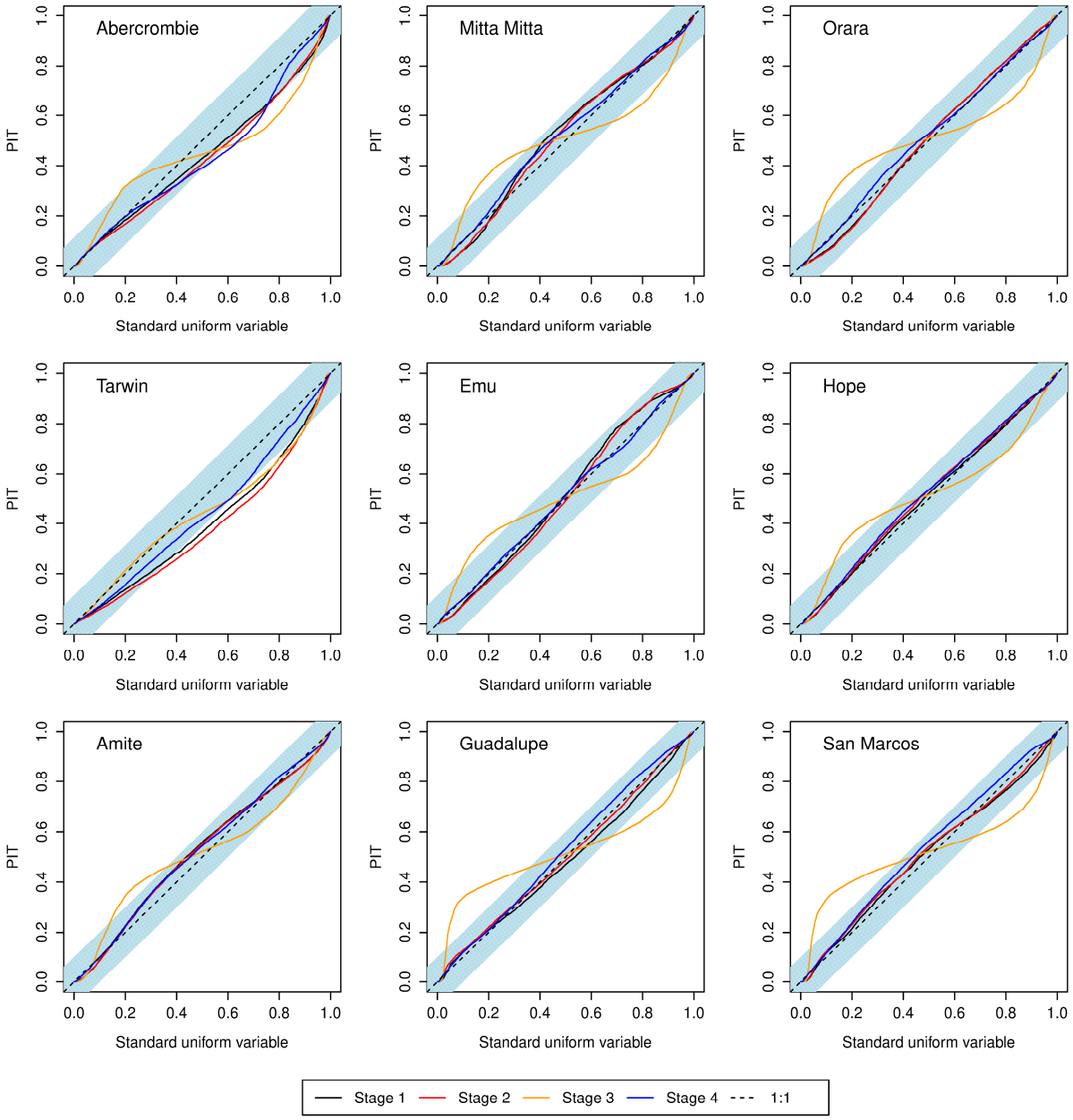
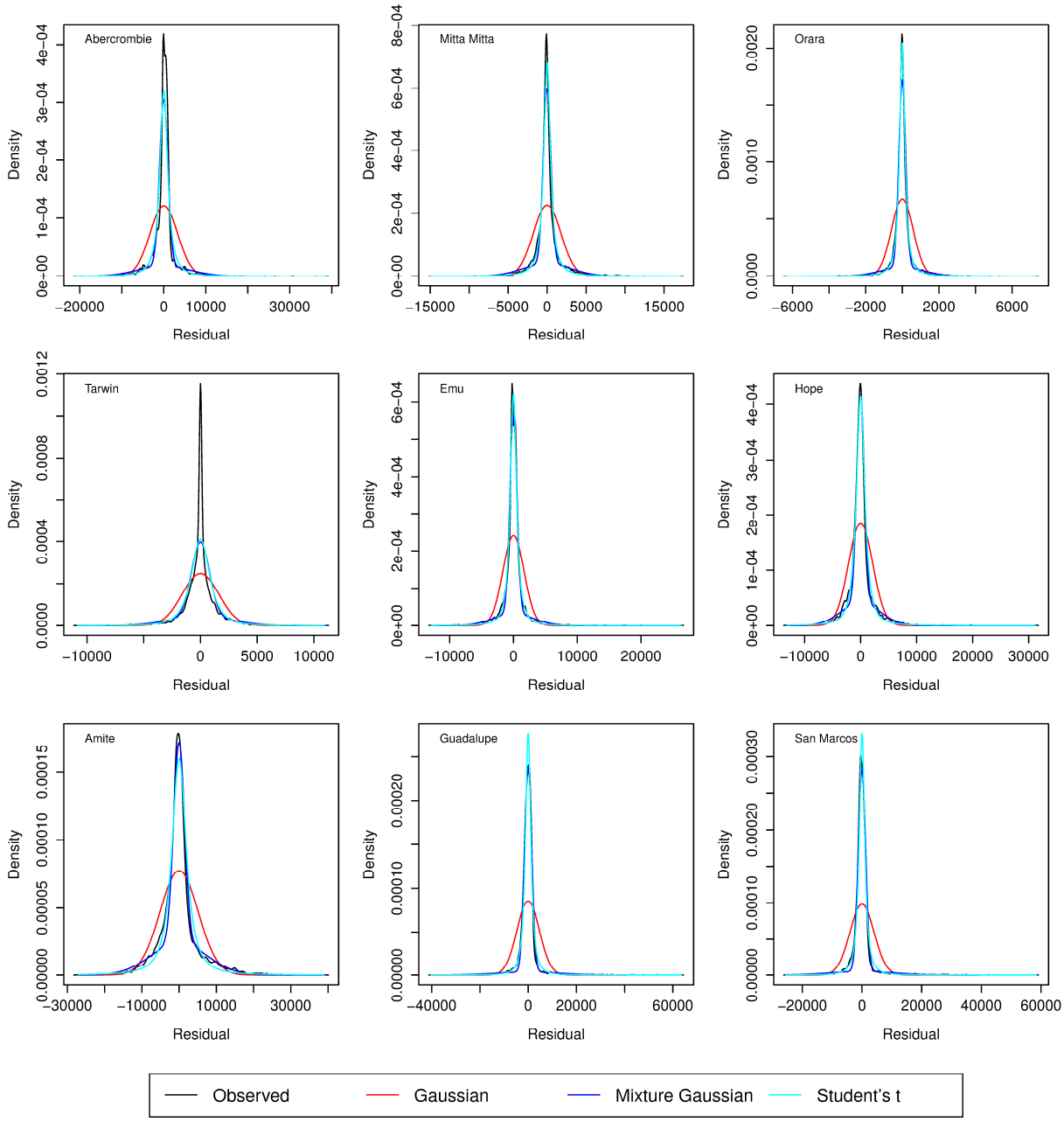


Figure 4: Comparison of the cumulative probability distribution of the PIT at different stages (light blue shaded strips indicate the 95% significant band of the Kolmogorov-Smirnov test.)



808

809 **Figure 5: Comparison of the different probability density functions fitted to the model residuals at Stage 3 for**
810 **each catchment.**
811

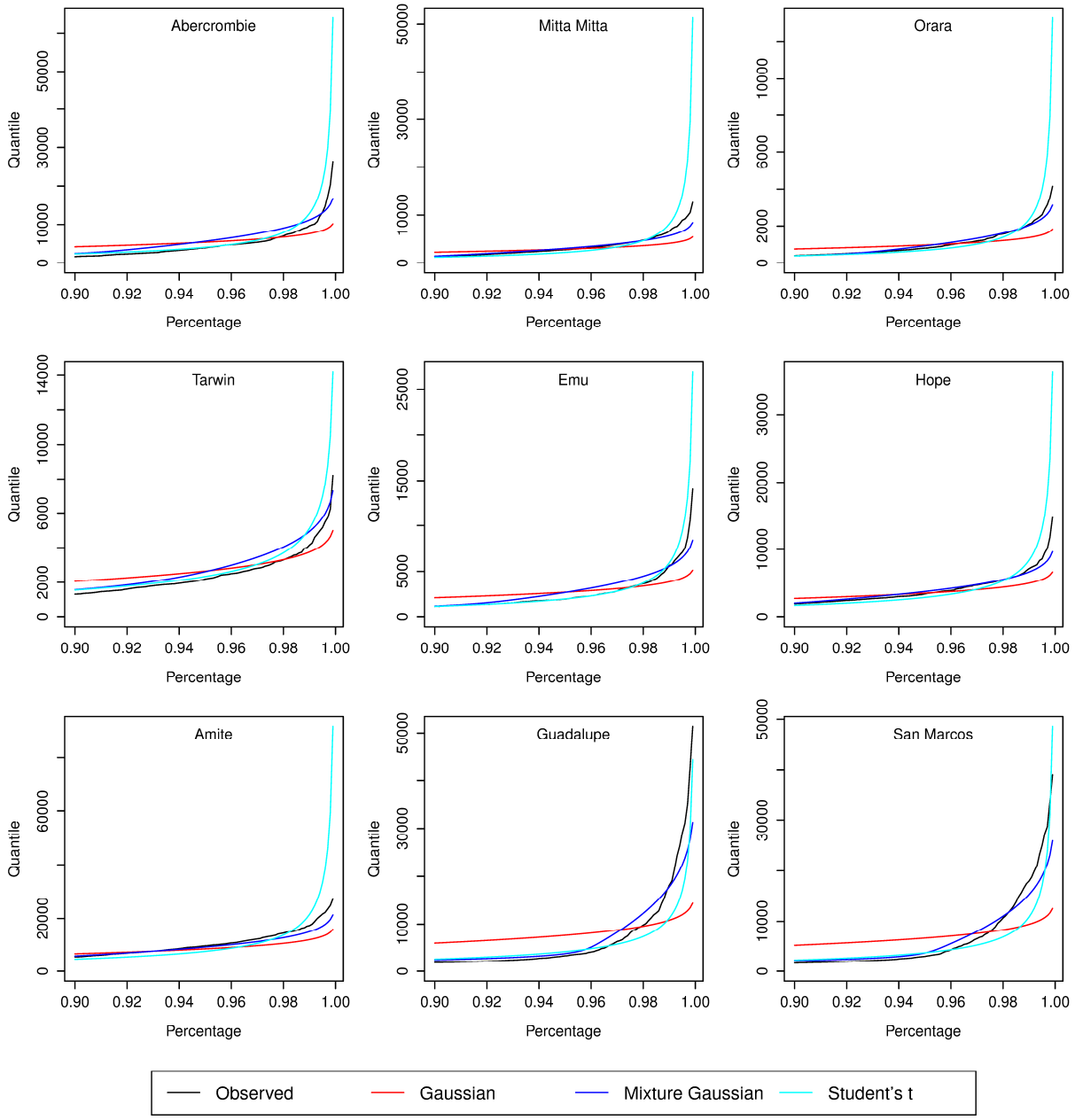
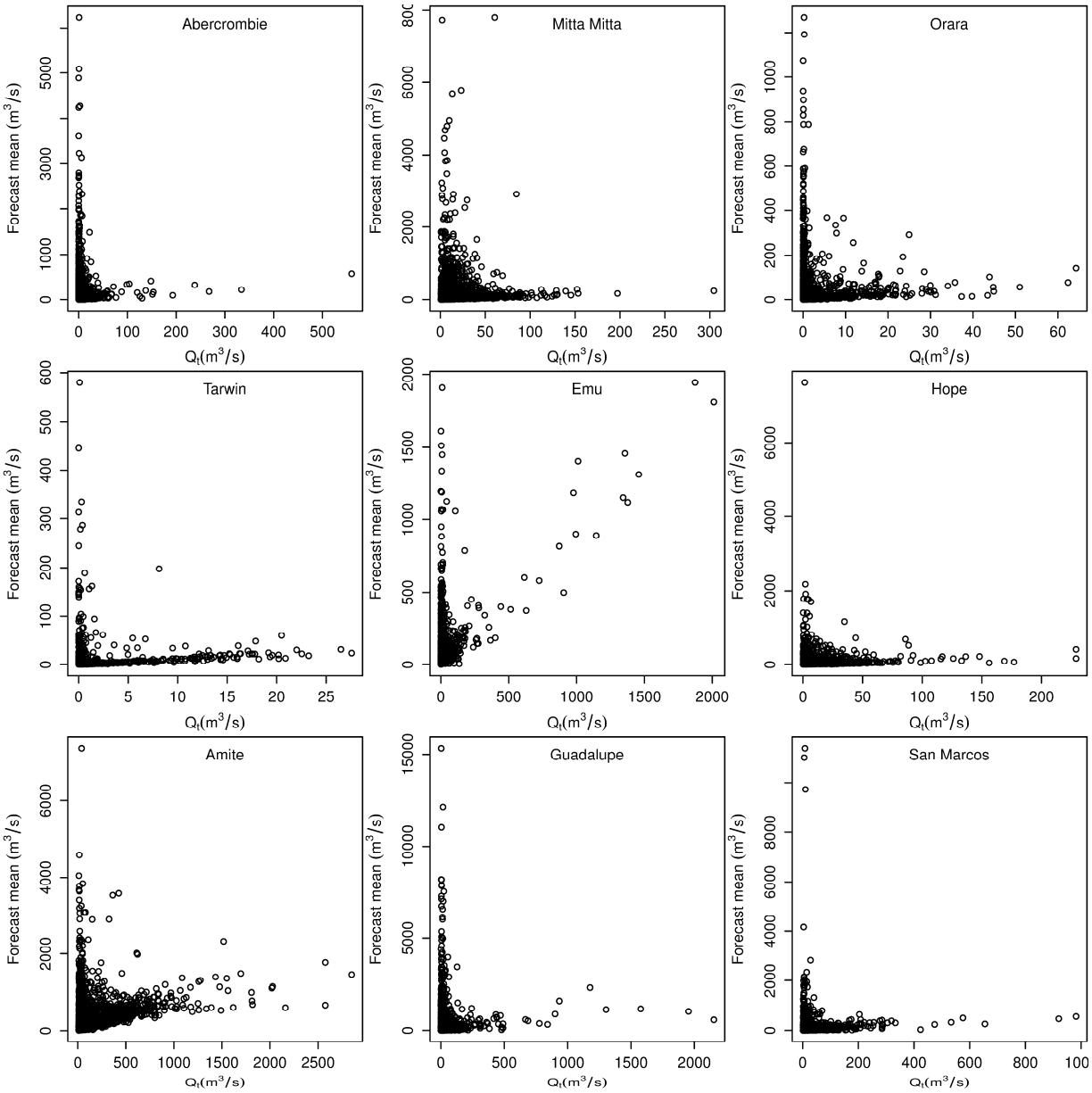


Figure 6: Comparison of the upper quantile of the model residuals fitted by different distributions for each catchment.

816
817



818
819
820 **Figure 7: Comparison of streamflow observations with streamflow forecast mean for each catchment when**
821 **the residual distribution is fitted by Student's t-distribution.**
822

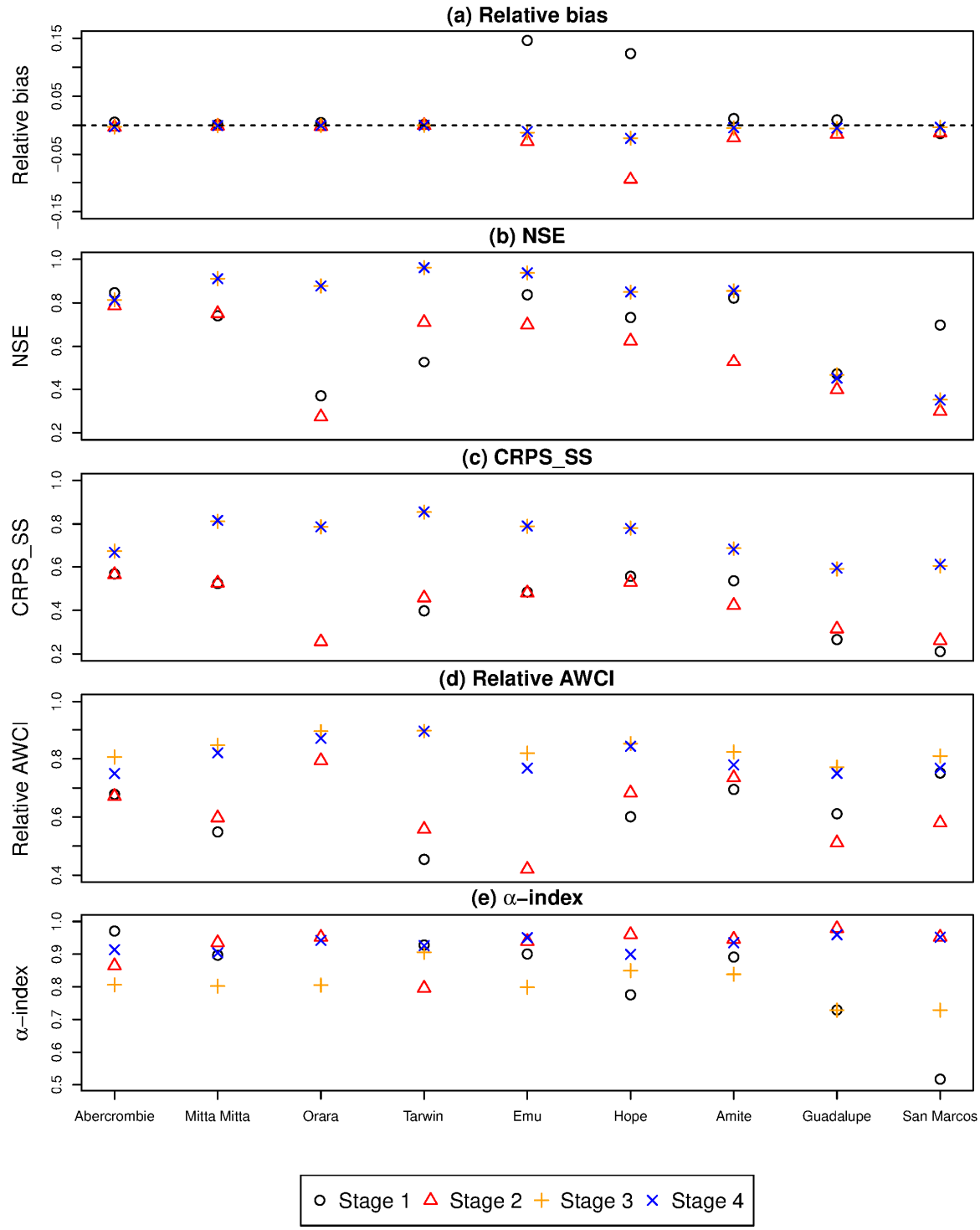
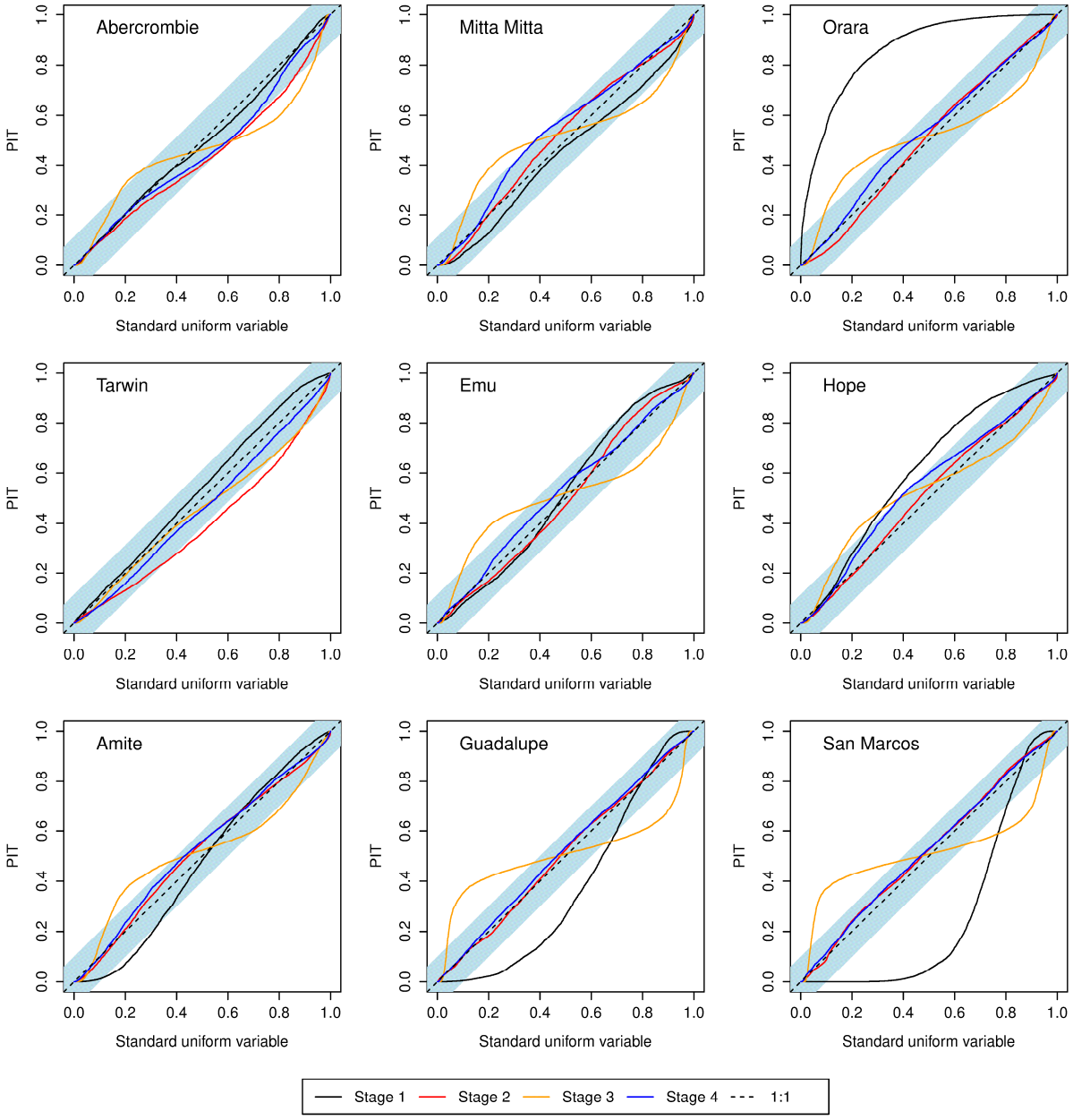


Figure 8: Same as Figure 3 but the hydrological model is calibrated by the least-squares method.

826
827



828
829 **Figure 9: Same as Figure 4 but the hydrological model is calibrated by the least-squares method.**
830

831

832 **Table 1: Summary of the ERRIS method**

	Stage 1	*Stage 2*	*Stage 3*	*Stage 4*
Purpose	Transformation and Hydrological model simulation	Linear bias correction	AR updating	Residual distribution refinement
Calibrated parameters	Hydrological model parameters, transformation parameters	bias-correction parameter	AR parameters	Distribution parameters
Correlation structure	Independent	Independent	Auto-correlated with lag one	Auto-correlated with lag one
Residual distribution	Transformed-Gaussian	Transformed -Gaussian	Transformed-Gaussian	Transformed- Gaussian mixture

833

834

835
836

837 **Table 2: Basic catchment characteristics (1992-2005)**

Name	Country	Gauge Site	Area (km^2)	Rainfall (mm/yr)	Streamflow (mm/yr)	Runoff coefficient	Zero flows
Abercrombie	Aus	Abercrombie River at Hadley no. 2	1447	783	63	0.08	14.4%
Mitta Mitta	Aus	Mitta Mitta River at Hinnomunjie	1527	1283	261	0.20	0
Orara	Aus	Orara River at Bawden Bridge	1868	1176	243	0.21	0.6%
Tarwin	Aus	Tarwin River at Meeniyan	1066	1042	202	0.19	0
Emu	Aus	Mount Emu Creek at Skipton	1204	641	23	0.04	0
Hope	Aus	Mount Hope Creek at Mitiamo	1646	436	11	0.02	23.3%
Amite	US	07378500	3315	1575	554	0.35	0
Guadalupe	US	08167500	3406	772	104	0.13	1.7%
San Marcos	US	08172000	2170	844	165	0.20	0

838

839

840

841 **Table 3: AWCI and CRPS calculated from the reference forecast for each catchment**

	Abercrombie	Mitta Mitta	Emu	Hope	Orara	Tarwin	Amite	Guadalupe	San Marcos
$AWCI_{REF}$ (m^3/s)	18.00	49.68	9.41	5.04	62.83	38.81	409.63	70.25	59.69
$CRPS_{REF}$ (m^3/s)	2.20	6.42	0.79	0.46	10.25	4.65	41.69	9.29	7.64

842

843 **Table 4: The calibrated error model parameters for the selected catchments.**

Stage	Parameter	Catchment								
		Abercrombie	Mitta Mitta	Emu	Hope	Orara	Tarwin	Amite	Guadalupe	San Marcos
1	x_1	551.26	1319.05	485.73	561.36	481.28	672.24	1279.63	763.15	906.72
	x_2	-0.41	-3.13	-3.22	-0.06	0.49	-2.20	-2.59	0.92	1.66
	x_3	7.94	65.63	12.40	1.10	28.71	20.24	44.67	23.67	39.93
	x_4	12.29	9.39	25.86	89.21	20.33	27.54	15.59	8.80	11.76
	$\log(a)$	-10.55	-9.70	-14.95	-11.80	-9.08	-11.55	-21.48	-10.38	-23.75
	$\log(b)$	-9.46	-9.49	-7.51	-8.68	-9.01	-9.35	-9.95	-9.89	-9.89
	σ_1	5298.92	5233.01	1790.99	4523.05	4490.65	5271.08	8885.27	8366.75	6843.48
2	c	6997.90	-14341.19	-373.84	946.83	-3153.26	-3282.81	1117.29	24909.80	10653.89
	d	1.06	0.85	0.98	1.02	0.95	0.96	1.01	1.16	1.07
	σ_2	5290.04	4924.38	1789.96	4540.44	4468.17	5244.14	8884.12	8025.35	6767.15
3	ρ	0.86	0.95	0.96	0.97	0.95	0.94	0.86	0.83	0.82
	σ_3	3289.50	1765.58	592.12	1611.67	1656.96	2154.72	5155.51	4661.31	4058.23
4	w	0.73	0.69	0.77	0.70	0.75	0.64	0.55	0.86	0.87
	S_1	1006.22	492.91	186.56	792.99	558.05	678.15	1481.79	1417.63	1246.49

	S_2	6238.76	3092.35	1192.76	2693.45	3159.56	3473.87	7487.62	9573.92	10673.07

844

845 **Table 5: The calibrated parameters when Student's t distribution is used to describe the residual distribution**
846 **at Stage 4**

	Abercrombie	Mitta Mitta	Emu	Hope	Orara	Tarwin	Amite	Guadalupe	San Marcos
r	1058.36	487.30	163.52	875.77	547.63	824.62	2033.78	1148.71	836.18
ν	1.44	1.25	1.33	2.31	1.53	1.58	1.62	1.36	1.54

847