



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

Ming Li¹, Q. J. Wang², James C. Bennett², and David E. Robertson²

¹CSIRO Data61, Floreat, WA, Australia

²CSIRO Land and Water, Clayton, VIC, Australia

Correspondence to: Ming Li (ming.li@csiro.au)

Received: 27 November 2015 – Published in Hydrol. Earth Syst. Sci. Discuss.: 20 January 2016

Revised: 4 July 2016 – Accepted: 5 August 2016 – Published: 6 September 2016

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.

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, using 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 the 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, performing parameter sensitivity analyses and comparing different hydrological models, without confounding input and output errors. However, because of a lack of information on the different sources of errors and on

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

how they interact with each other, it is highly challenging to apply an error decomposition approach to arrive at statistically reliable overall errors in streamflow forecasts (Renard et al., 2010).

An alternative approach is to consider only the overall errors of forecasts, without attempting to explain the sources of errors. An estimate of the overall error of a forecast is the residual, defined as the difference between modelled streamflow and observations. We now concentrate our discussion on residuals, but we will continue to refer to models of residuals as “error models”, following common practice. Residuals of a series of forecasts form a time series. The most traditional and simplest error model, related to the classical least-squares calibration, is based on the assumption of uncorrelated homoscedastic Gaussian residuals in the time series of residuals (Diskin and Simon, 1977). This assumption is generally not valid for hydrological applications, where residuals are frequently autocorrelated, heteroscedastic and non-Gaussian (Kuczera, 1983; Sorooshian and Dracup, 1980). More sophisticated error models have been developed to address correlation, variance structure and the distribution of residuals. Autoregressive models have been widely used to account for autocorrelation of residuals (e.g. Bates and Campbell, 2001; Xiong and O’Connor, 2002). Heteroscedasticity may be explicitly dealt with by describing the variance of residuals as a function of some state-dependent variables (e.g. observed streamflow, dry/wet seasons) (e.g. Evin et al., 2013; Pianosi and Raso, 2012; Schaeffli et al., 2007; Yang et al., 2007). Non-Gaussianity of residuals may be explicitly represented by non-Gaussian probability distributions (e.g. Marshall et al., 2006; Schaeffli et al., 2007; Schoups and 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 behaviour of errors, including behaviours that arise from interactions be-

tween 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 autocorrelated 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. Gagne 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 Sect. 2. A case study is introduced in Sect. 3. Major results are presented in Sect. 4, followed by discussion and further results in Sect. 5. Conclusions are made in Sect. 6.

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

2.1 Model formulation

2.1.1 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)

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

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

$$Z(t) = f(Q(t)), \quad (2)$$

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

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

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.

2.1.2 Stage 2: linear bias correction

At Stage 1, we assume that the hydrological simulation is overall unbiased. However, the hydrological model often overestimates low flows and underestimates high flows. At Stage 2, we adopt a simple but effective bias-correction scheme firstly introduced by Wang et al. (2014) to revise 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

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

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

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 addition 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.

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	Autocorrelated with lag one	Autocorrelated with lag-1
Residual distribution	Transformed-Gaussian	Transformed-Gaussian	Transformed-Gaussian	Transformed-Gaussian mixture

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

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

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

2.1.3 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

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

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

where $\tilde{Q}_3^*(t) = f^{-1}(\tilde{Z}_2(t) + \rho(Z(t-1) - \tilde{Z}_2(t-1)))$ 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, overcorrection 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 overcorrect forecasts. In Eq. (8) the forecast median, denoted by $\tilde{Q}_3(t)$, is given by

$$\tilde{Q}_3(t) =$$

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

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

2.1.4 Stage 4: residual distribution refinement

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

At Stage 4, we use a non-Gaussian distribution to describe the model residuals from Stage 3. Several long-tailed distributions have been used in hydrological modelling studies, such as the finite mixture distribution (Schaeffli et al., 2007; Smith et al., 2010), the exponential power distribution (Schoups and Vrugt, 2010) and Student's t distribution (Marshall et al., 2006). In this study, we assume that the model residuals can be grouped into two categories with respect to variance and thus choose a two-component Gaussian mixture distribution. It is 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 Sect. 5.1.

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

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

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

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 $\left(Z(t)|\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right)$, can be explicitly written as a weighted sum of two Gaussian probability density functions:

$$\begin{aligned} \text{PDF}\left(Z(t)|\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right) &= w\phi\left(Z(t)|\tilde{Z}_4(t), \sigma_{4,1}^2\right) \\ &+ (1-w)\phi\left(Z(t)|\tilde{Z}_4(t), \sigma_{4,2}^2\right), \end{aligned} \quad (13)$$

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

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

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

$$L_4(\theta_4) = \prod_t J_{z \rightarrow Q} \text{PDF}\left(Z(t)|\tilde{Z}_4(t), \sigma_{4,1}^2, \sigma_{4,2}^2, w\right), \quad (15)$$

where $J_{z \rightarrow Q} = 1/\tanh\{a + bQ(t)\}$ 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 Sect. 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 autocorrelation coefficient (ρ) and the residual

standard deviation (σ_3) are estimated. At the end of Stage 3, the value of the autocorrelation 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

$$\pi_t = F_t(Q(t)), \quad (16)$$

where F_t is the forecast cumulative distribution function 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

$$\alpha = 1 - \frac{2}{n} \sum_{i=1}^n \left| \pi_i^* - \frac{i}{n+1} \right|, \quad (17)$$

where π_i^* 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:

$$\text{Relative AWCI} = \frac{\text{AWCI}_{\text{REF}} - \text{AWCI}}{\text{AWCI}_{\text{REF}}}, \quad (18)$$

where AWCI_{REF} is AWCI calculated from the reference forecast. The reference forecast in this study is generated by re-sampling 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 14 years 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 1, 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; Gritti 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:

$$\text{CRPS}_{\text{SS}} = \frac{\text{CRPS}_{\text{REF}} - \text{CRPS}}{\text{CRPS}_{\text{REF}}}, \quad (19)$$

where CRPS_{REF} is CRPS calculated from the reference forecast (already defined for Eq. 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 southeastern Australia and three catchments in the United States (US) for this study (Fig. 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) data set (Vaze et al., 2011). The rainfall and potential evaporation data for the Australian catchments are taken from the Australian Water Availability Project (AWAP) data set (Jones et al., 2009). All data for the US catchments are taken from the Model Intercomparison Experiment (MOPEX) data set (Duan et al., 2006). The Abercrombie and Emu catchments have many instances of zero flow (Table 2), and accurate streamflow forecasting is particularly challenging for such dry catchments. AWCI_{REF} and CRPS_{REF} for each catchment are given in Table 3.

3.2 Cross-validation

Daily streamflow is simulated with the GR4J rainfall–runoff model (Perrin et al., 2003) and then forecasted with ERRIS as described in Sect. 3. GR4J is a widely used conceptual model that was designed to be as parsimonious as possible. Its four parameters describe the depth of a production store (X1), groundwater exchange (X2), the depth of a routing store (X3) and the length of unit hydrographs (X4). Forecasts are generated from “perfect” (observed) deterministic rainfall forecasts at a lead time of 1 day (i.e. one time step ahead). All results reported in this study are based on cross-validation unless specified. Cross-validation allows us to generalize the forecast skill to data outside the sample period. Because of data availability, we choose different study periods for Australian and US catchments. For Australian catchments, data from 1990 to 1991 are used to warm up the hydrological model and the data from 1992 to 2005 are used to generate a leave-two-years-out cross-validation (i.e. effectively 14-fold cross-validation). For a particular year, we remove the streamflow data from this year and the following year and apply ERRIS to forecast the streamflow for the year. The removal of the data from the following year aims to minimize the impact of streamflow memory on model performance. For US catchments, the data from 1979 to 1980 are used in the warm-up period and the data from 1981 to 1998 are used for a leave-two-years-out cross-validation (i.e. effectively 18-fold cross-validation).

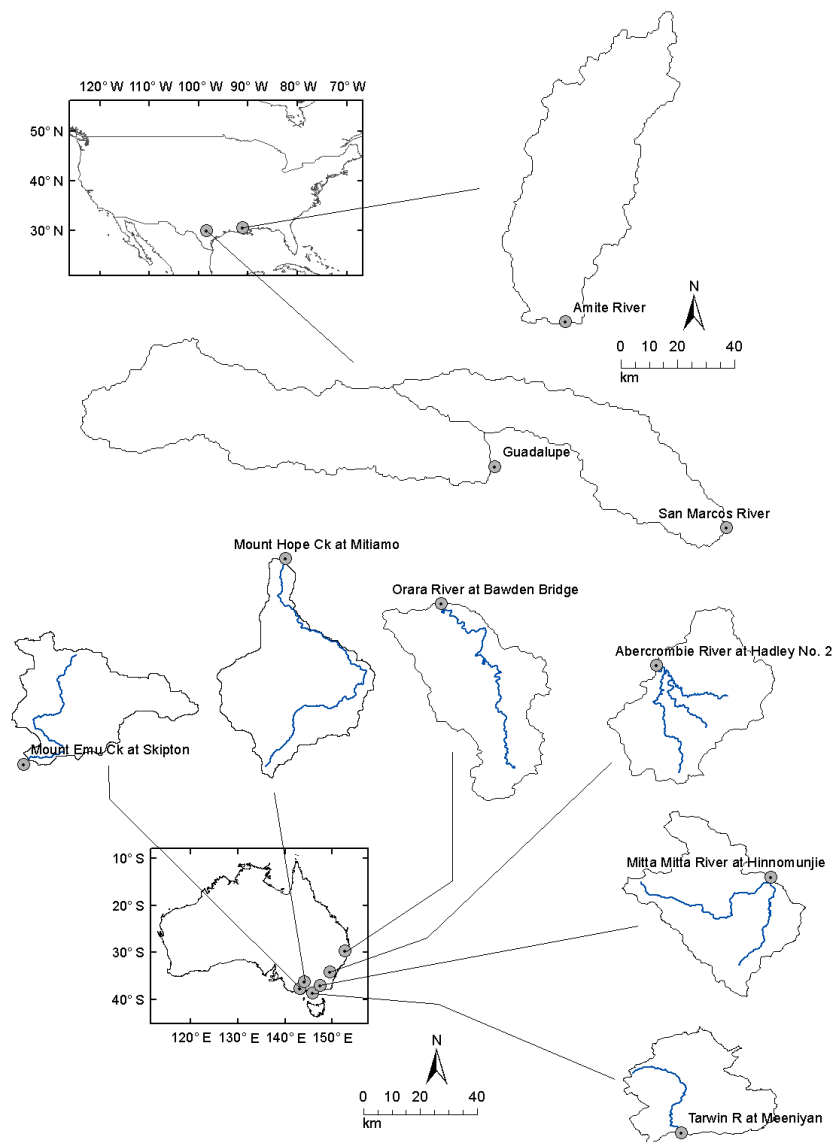


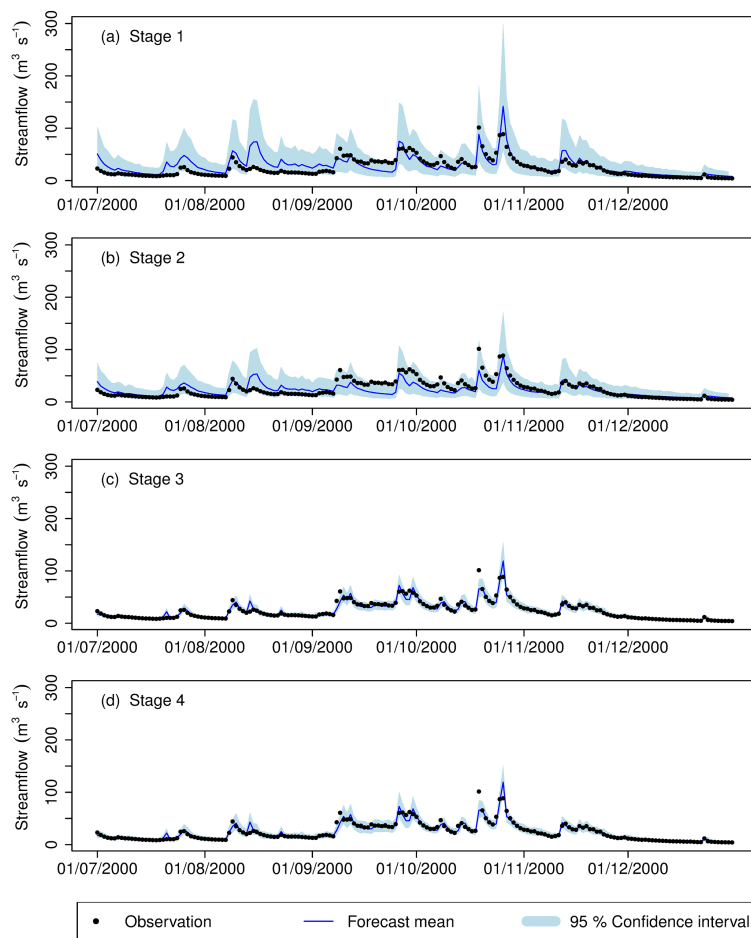
Figure 1. Map of the catchments used in this study.

Table 2. Basic catchment characteristics (1992–2005).

Name	Country	Gauge site	Area (km ²)	Rainfall (mm yr ⁻¹)	Streamflow (mm yr ⁻¹)	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

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 ⁻¹)	18.00	49.68	9.41	5.04	62.83	38.81	409.63	70.25	59.69
CRPS _{REF} (m ³ s ⁻¹)	2.20	6.42	0.79	0.46	10.25	4.65	41.69	9.29	7.64

**Figure 2.** An example of streamflow time-series plots for the Mitta Mitta catchment in the period between 1 July 2000 and 31 December 2000.

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 1 July 2000 and 31 December 2000. The forecast mean and the 95 % forecast interval are plotted against observations. The forecast at Stage 1 (the base hydrological model forecast) frequently overestimates 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

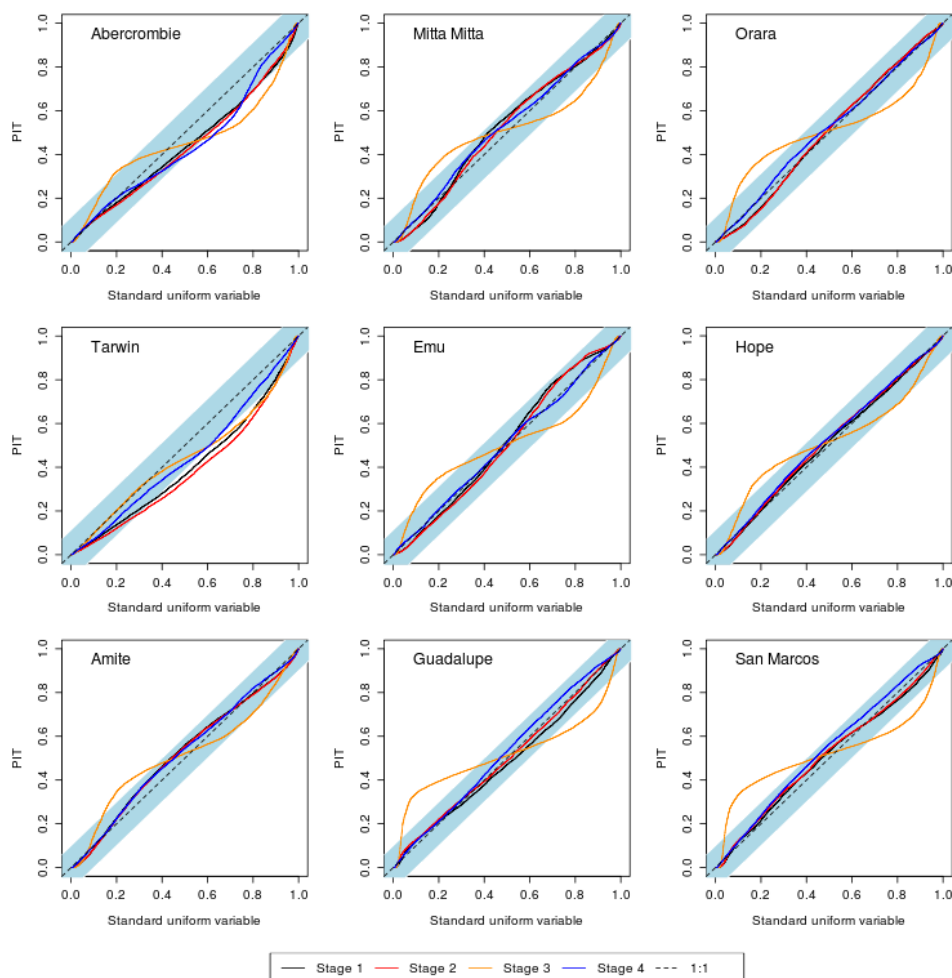


Figure 3. Comparison of performance metrics for each catchment and each stage.

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 Fig. 2. In general, Stage 1 and Stage 2 are similarly efficient (Fig. 3b), skillful (Fig. 3c), sharp (Fig. 3d) and reliable (Fig. 3e). As we expect, Stage 2 forecasts are consistently less biased than Stage 1 (Fig. 3a) (except for the Hope catchment, where many instances of zero flow occur; see Table 2). Stage 3 is generally much more efficient and skillful than Stages 1 and 2. A partial exception to this is the Abercrombie catchment, which is less efficient at Stage 3 than Stage 2. The Abercrombie catchment experiences low (to zero) flows, but is also punctuated by abrupt high flows. Stage 3 is based on the time persistence of the residuals and may introduce more errors when flows change abruptly, which sometimes occurs in the Abercrombie catchment. In addition, residuals tend to be larger at higher flows and because NSE is a measure of squared residuals, it tends to give more weights to residuals at high flows. This causes

the Abercrombie Stage 3 forecasts to be less efficient than those of Stage 2.

As we expect, Stage 3 forecasts are notably sharper than those at Stage 2 (Fig. 3d). However, this sharpness is not supported by reliability: Stage 3 forecasts tend to be much less reliable than all other stages (Fig. 3e). Figure 4 illustrates the reliability of the forecasts at each stage in more detail with the PIT plots. As PIT values are highly autocorrelated, we have to “thin” them in order to make the Kolmogorov–Smirnov significant band applicable (Zhao et al., 2015). We generate PIT plots from every 30th forecast to eliminate the autocorrelation. The PIT plots show that the forecasts at the first two stages are reliable (as with the α index in Fig. 3e). However, for Stage 3 the points on the PIT plots deviate substantially from the 1 : 1 line, with a clear S-shape pattern for almost all catchments (the exception is the Tarwin catchment). A traditional interpretation of this S shape is that the forecasts are underconfident (Laio and Tamea, 2007). However, in this case, the S shape is caused by the high level of kurtosis in the distribution of the residuals, as we will show

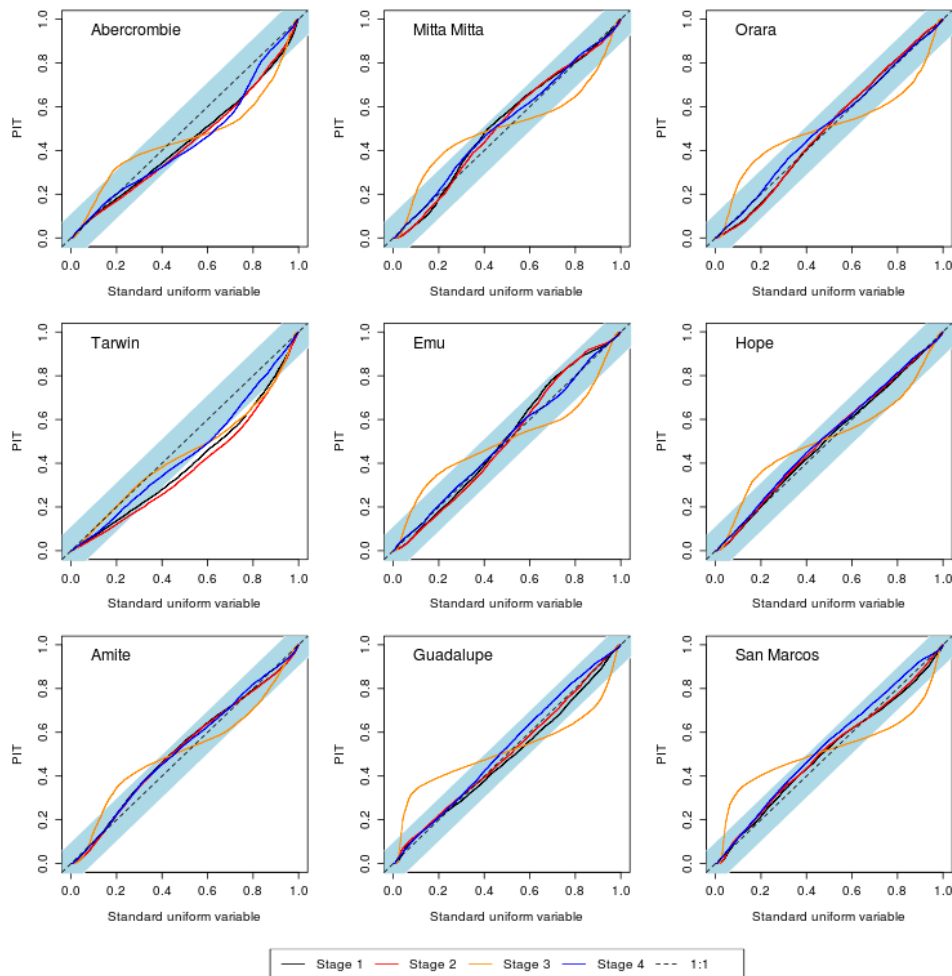


Figure 4. Comparison of the cumulative probability distribution of the PIT at different stages (light-blue shaded strips indicate the 95 % significant band of the Kolmogorov–Smirnov test).

below. The α index from Stage 3 is smaller than those indices from Stages 1 and 2 (the Tarwin catchment is the only exception), confirming the lack of the reliability at Stage 3. Stage 4 consistently improves the reliability of the forecast after the AR updating. The PIT plot at Stage 4 is much closer to the 1 : 1 line than that at Stage 3, and this is reflected by the α index, which increases for all catchments. Stage 4 corrects the underconfident forecasts from Stage 3 and slightly decreases the sharpness from Stage 3 (Fig. 3d).

At Stage 3, unreliable forecasts are caused by representing the model residual by an inappropriate (Gaussian) probability distribution. We compare the underlying density of the model residuals at Stage 3, $\varepsilon(t) = Z_3(t) - \bar{Z}_3(t)$ (fitted by the nonparametric density estimation), with the fitted parametric densities for different distributions in Fig. 5. The fitted Gaussian density is flatter than the underlying density of $\varepsilon(t)$ in order to match the tails for each catchment. This suggests that the residual distribution is more peaked and has longer tails than the Gaussian distribution. As we have seen above,

forecast residuals are, in general, dramatically reduced by the AR error updating. Unfortunately, this reduction in residuals does not occur at all events, especially where abrupt changes in flow occur (and hence the assumption of strong autocorrelation breaks down). Thus, the magnitude of the forecast residuals at Stage 3 for a small proportion of events is large relative to the majority of events. As we have seen, the practical implication of the dichotomous behaviour of the residuals is that their distribution is still bell-shaped and symmetric but is peakier and has a much longer tail than the Gaussian distribution. The Gaussian mixture distribution treats model residuals as two groups with different variances. The Gaussian mixture distribution is able to capture the peak and tails of the underlying residual density for all catchments, resulting in reliable ensemble forecasts that also have a highly accurate forecast mean. As we note in the introduction, however, other distributions have also been used to describe “peaky” data, and we explore these in the next section.

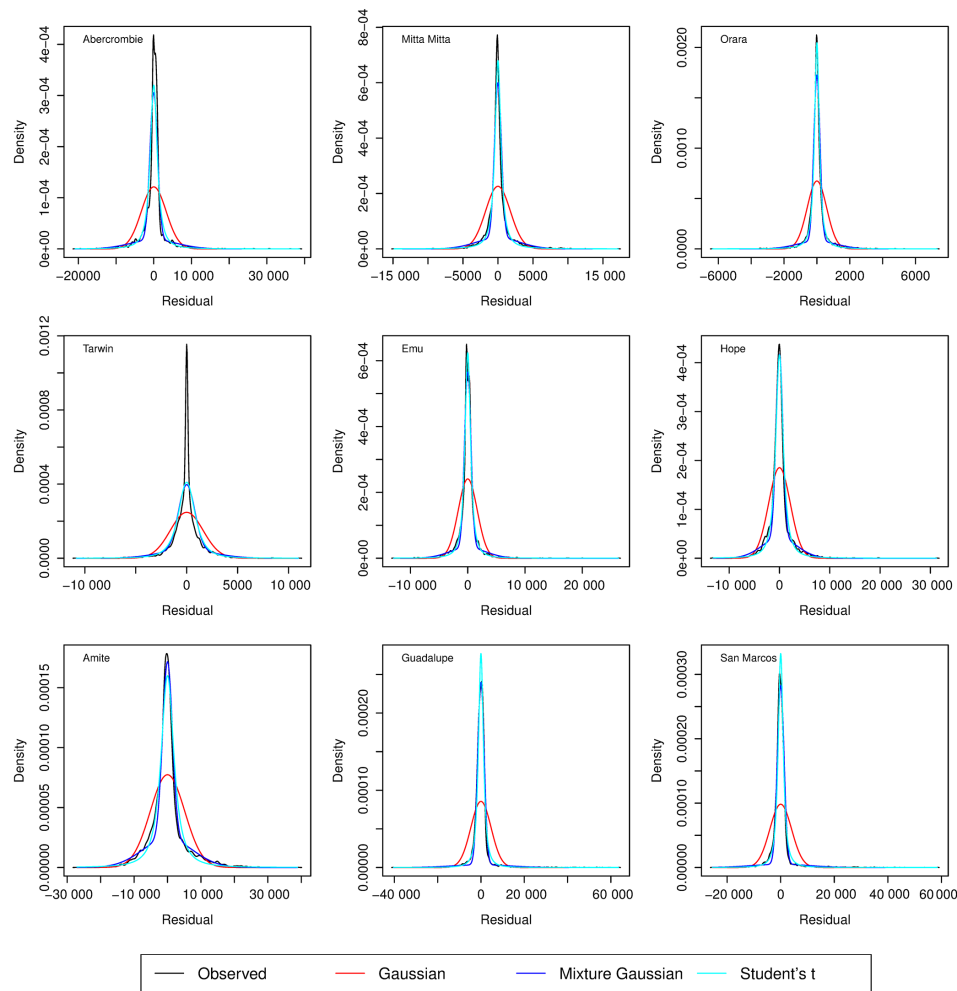


Figure 5. Comparison of the different probability density functions fitted to the model residuals at Stage 3 for each catchment.

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

5 Further results

5.1 Testing an alternative residual distribution

It is possible to use long-tailed distributions other than the Gaussian mixture distribution at Stage 4. For example, Student's t distribution is a simple long-tailed distribution that

has been used in hydrological modelling (e.g. Marshall et al., 2006). In this section we investigate whether Student's t distribution is a viable alternative to the Gaussian mixture distribution at Stage 4. To do this, we modify the model residual in Eq. (12) as follows:

$$Z(t) = \tilde{Z}_4(t) + r\xi(t), \quad (20)$$

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

We first examine how well Student's t distribution can fit the residual distribution at Stage 4 for all nine catchments (Fig. 5). High peaks and long tails of the residual densities can be captured reasonably well by Student's t distribution for nearly all catchments. The fitted densities of Student's t distribution appear more “peaked” for most catchments than those of the Gaussian mixture distribution, which is originally used at Stage 4. Figure 6 further investigates how Student's t distribution can fit the upper quantile of the model

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	-14 341.19	-373.84	946.83	-3153.26	-3282.81	1117.29	24 909.80	10 653.89
	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	10 673.07

residuals. There is a clear tendency of Student's t distribution to overestimate the upper quantile (e.g. 98 % or higher) of the model residuals (especially for the Australian catchments). These upper quantiles are more accurately estimated by the Gaussian mixture distribution. This implies that Student's t distribution often has tails that are too long. We note, however, that if the ERRIS method is tested on other catchments, it is possible that Student's t distribution may describe the residuals better than the Gaussian mixture distribution in some cases.

A disadvantage of the very long tail of Student's t distribution is that it can be problematic for operational forecasting. The degrees of freedom, ν , determine how heavy the tails of Student's t distribution are. Table 5 presents the two calibrated parameters (i.e. ν and r) for all catchments. Calibrated ν values are less than 2 for eight out of nine catchments. The exception is the Hope catchment, and even here the calibrated ν is very close to 2. It is well known that for degrees of freedom less than 2, Student's t distribution is so heavy-tailed that the variance is infinite (if $1 < \nu \leq 2$) or even undefined (if $\nu \leq 1$). This is obviously undesirable for operational forecasting: it can cause a few forecast ensemble members to be so large that the forecast mean becomes implausibly large. Figure 7 compares the forecast mean with observations if the model residual is revised as Eq. (19). In all catchments, in some cases forecast mean values are unrealistically large even as observations are relatively small. Student's t distribution is thus prone to being 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 Fig. 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 Stages 3 and 4. Thus, ERRIS is flexible enough to accommodate existing hydrological models.

Figure 9, an analogue to Fig. 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 are no longer reliable in many instances. This is caused by the least-squares calibration, which does not ensure the forecast residuals are Gaussian (even after the log-sinh transformation). The PIT plots derived from Stage 2 and

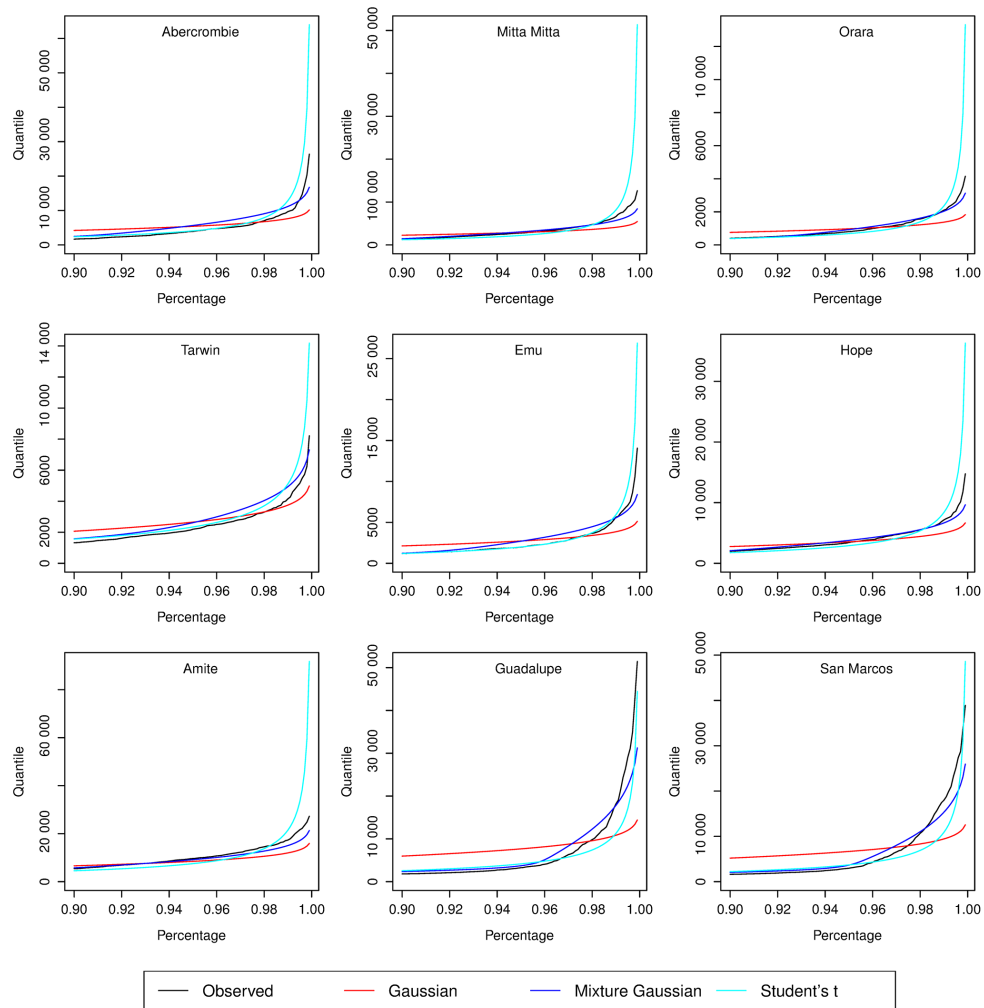


Figure 6. Comparison of the upper quantile of the model residuals fitted by different distributions for each catchment.

Table 5. The calibrated parameters when Student’s *t* distribution is used to describe the residual distribution at Stage 4.

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

Stage 3 in Fig. 9 show a very similar pattern to their counterparts in Figure 4. It suggests that poor reliability at Stage 3 occurs irrespective of the calibration strategy employed for the hydrological model. As with Fig. 4, Fig. 9 shows the Gaussian mixture distribution used at Stage 4 effectively ameliorates the problems with the reliability of Stage 3.

6 Discussion

There are several advantages of using a multi-stage error model compared to a single complex error model. (1) The parameter estimation in ERRIS is relatively simple, and hence

computationally efficient. Only a small number of parameters are estimated at each stage. Joint parameter estimations associated with a single complicated error model are often more computationally demanding. (2) Interference between parameters is minimized. The parameters of a single complex model can confound each other and the contribution of one parameter can sometimes be explained by others. For example, the hydrological model parameters describing soil moisture storage capacity may interfere strongly with the error parameters describing bias. Interference between parameters can make the parameter estimation unstable, because more than one set of parameters can achieve a similar ob-

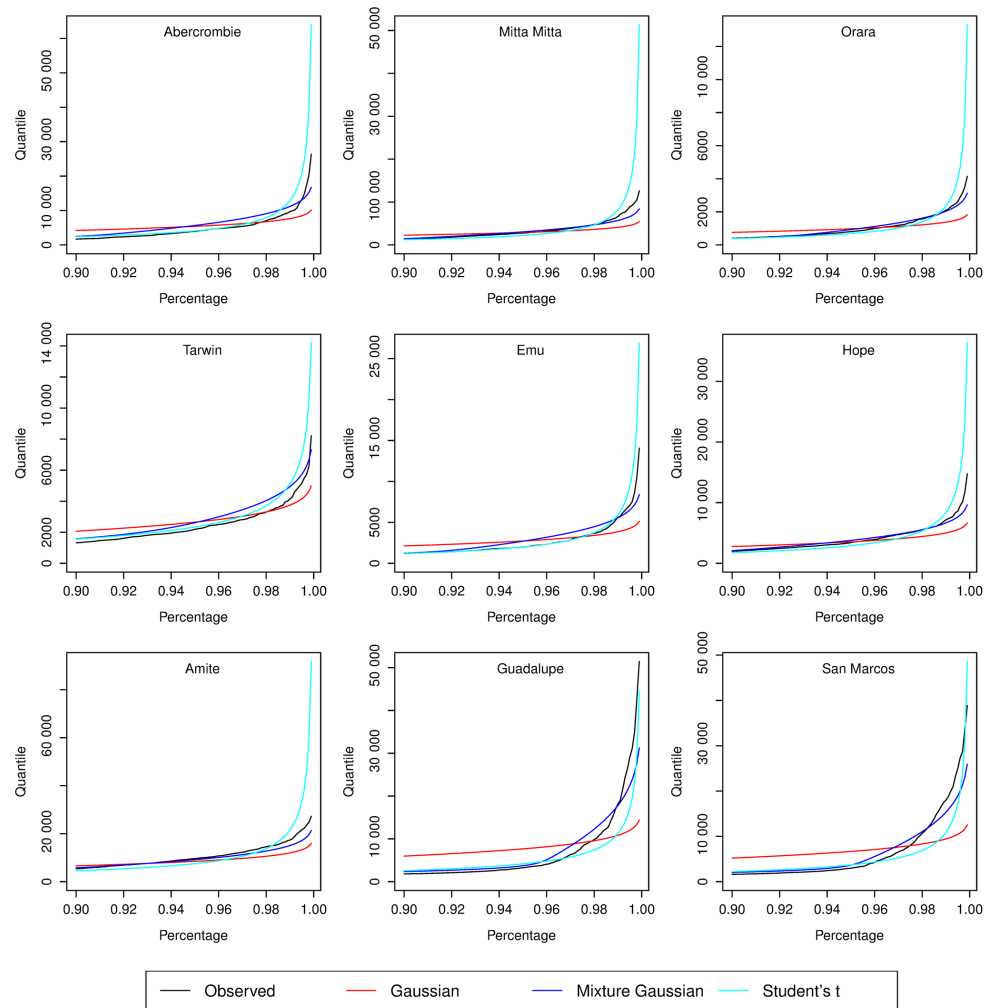


Figure 7. Comparison of streamflow observations with streamflow forecast mean for each catchment when the residual distribution is fitted by Student's t distribution.

jective function value, and thus over-fit parameters. (3) In operational forecasting it is often important that individual components of the forecasting model can function independently. For example, if forecasts are issued to long lead times, the influence of an AR model diminishes as lead time extends. Thus forecasts at long lead 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 1 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

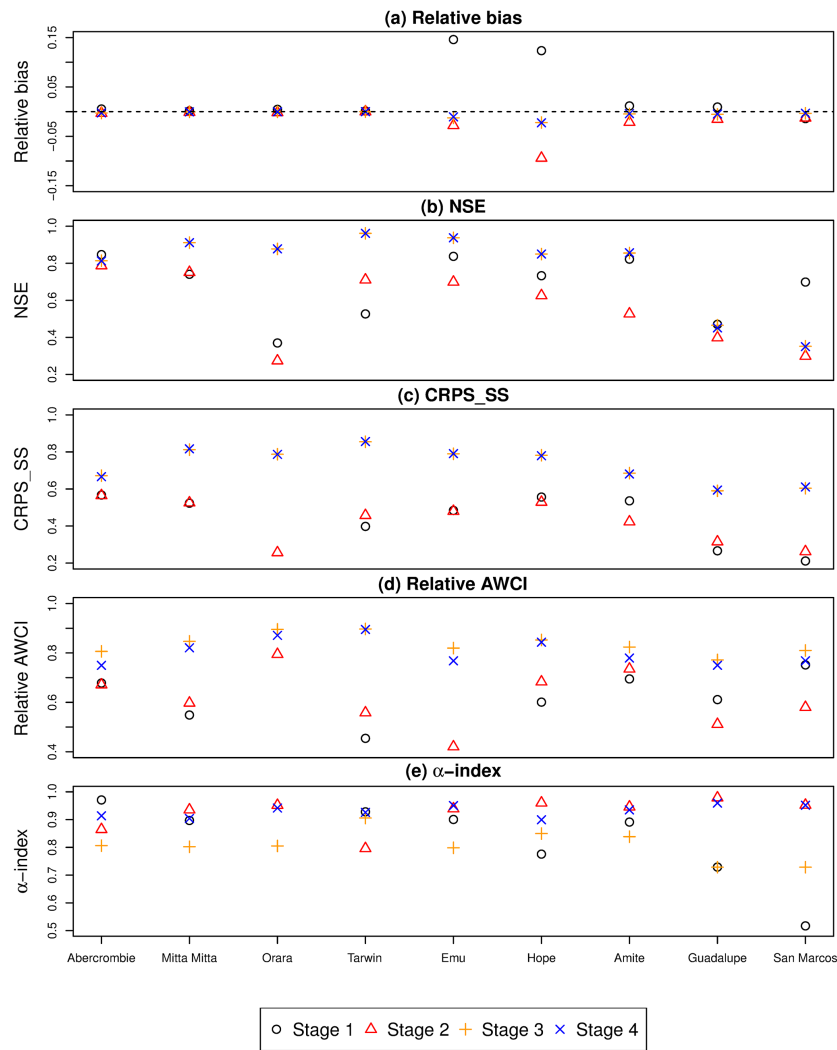


Figure 8. Same as Fig. 3 but the hydrological model is calibrated by the least-squares method.

correction, and (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 interaction between parameters, and ensures 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 DUMBRAE 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

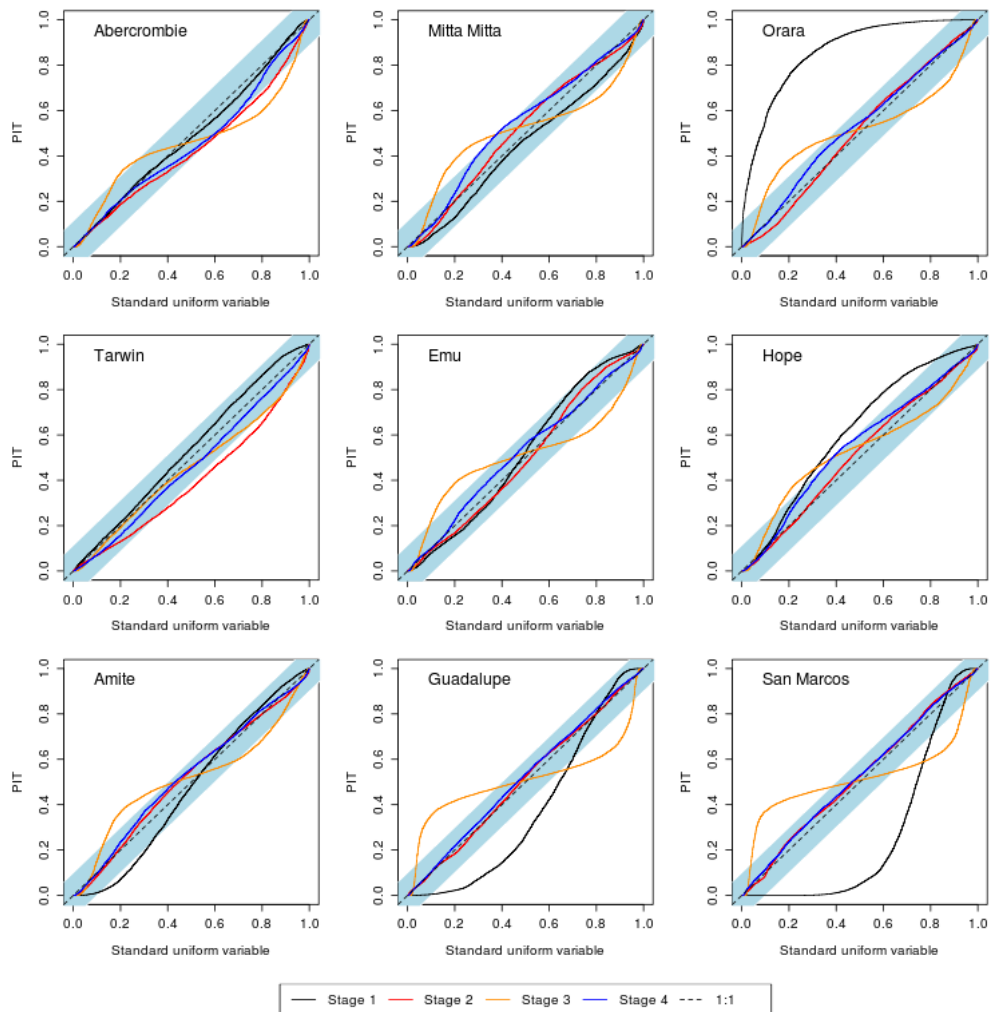


Figure 9. Same as Fig. 4 but the hydrological model is calibrated by the least-squares method.

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 shown 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.

8 Data availability

All data for the US catchments are taken from the MOPEX data set, which are available at http://www.nws.noaa.gov/ohd/mopex/mo_datasets.htm. All data for the Australian catchments are available from the corresponding author, Ming Li (ming.li@csiro.au), upon request, because the underlying data sets CWYET and AWAP (see Sect. 3.1) are not publicly available.

Acknowledgements. This research was 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.

Edited by: D. Solomatine

Reviewed by: two anonymous referees

References

- Ajami, N. K., Duan, Q. Y., and Sorooshian, S.: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, 43, W01403, doi:10.1029/2005wr004745, 2007.
- Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS – global ensemble streamflow forecasting and flood early warning, *Hydrol. Earth Syst. Sci.*, 17, 1161–1175, doi:10.5194/hess-17-1161-2013, 2013.
- Bates, B. C. and Campbell, E. P.: A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resour. Res.*, 37, 937–947, doi:10.1029/2000wr900363, 2001.
- Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q. J., Enever, D., Hapuarachchi, P., and Tuteja, N. K.: A System for Continuous Hydrological Ensemble to lead times of 9 days Forecasting (SCHEF), *J. Hydrol.*, 519, 2832–2846, doi:10.1016/j.jhydrol.2014.08.010, 2014a.
- Bennett, J. C., Wang, Q. J., Pokhrel, P., and Robertson, D. E.: The challenge of forecasting high streamflows 1–3 months in advance with lagged climate indices in southeast Australia, *Nat. Hazards Earth Syst. Sci.*, 14, 219–233, doi:10.5194/nhess-14-219-2014, 2014b.
- Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, *Q. J. Roy. Meteor. Soc.*, 131, 2131–2150, doi:10.1256/qj.04.71, 2005.
- Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrol. Earth Syst. Sci.*, 17, 4209–4225, doi:10.5194/hess-17-4209-2013, 2013.
- Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, *B. Am. Meteorol. Soc.*, 95, 79–98, doi:10.1175/bams-d-12-00081.1, 2014.
- Diskin, M. H. and Simon, E.: A procedure for the selection of objective functions for hydrologic simulation models, *J. Hydrol.*, 34, 129–149, doi:10.1016/0022-1694(77)90066-X, 1977.
- Duan, Q. Y., Sorooshian, S., and Gupta, V. K.: Optimal Use of the Sce-Ua Global Optimization Method for Calibrating Watershed Models, *J. Hydrol.*, 158, 265–284, doi:10.1016/0022-1694(94)90057-4, 1994.
- Duan, Q. Y., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, 320, 3–17, doi:10.1016/j.jhydrol.2005.07.031, 2006.
- Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model, *J. Hydrol.*, 384, 142–155, doi:10.1016/j.jhydrol.2010.01.018, 2010.
- Evin, G., Kavetski, D., Thyer, M., and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water Resour. Res.*, 49, 4518–4524, doi:10.1002/wrcr.20284, 2013.
- Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, 50, 2350–2375, doi:10.1002/2013WR014185, 2014.
- Gneiting, T. and Katzfuss, M.: Probabilistic Forecasting, *Annu Rev. Stat. Appl.*, 1, 125–151, doi:10.1146/annurev-statistics-062713-085831, 2014.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, 133, 1098–1118, doi:10.1175/Mwr2904.1, 2005.
- Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, *J. Roy. Stat. Soc. B*, 69, 243–268, doi:10.1111/j.1467-9868.2007.00587.x, 2007.
- Gragne, A. S., Sharma, A., Mehrotra, R., and Alfredsen, K.: Improving real-time inflow forecasting into hydropower reservoirs through a complementary modelling framework, *Hydrol. Earth Syst. Sci.*, 19, 3695–3714, doi:10.5194/hess-19-3695-2015, 2015.
- Grimm, E. P., Gneiting, T., Berrocal, V. J., and Johnson, N. A.: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification, *Q. J. Roy. Meteor. Soc.*, 132, 2925–2942, doi:10.1256/qj.05.235, 2006.
- Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, 129, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.
- Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecast.*, 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.
- Jones, D. A., Wang, W., and Fawcett, R.: High-quality spatial climate data-sets for Australia, *Australian Meteorological and Oceanographic Journal*, 58, 233–248, 2009.
- Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, 42, W03407, doi:10.1029/2005wr004368, 2006a.
- Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, 42, W03408, doi:10.1029/2005wr004376, 2006b.
- Kelly, K. S. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, *Stoch. Hydrol. Hydraul.*, 11, 17–31, doi:10.1007/Bf02428423, 1997.

- Krzysztofowicz, R.: Transformation and normalization of variates with specified distributions, *J. Hydrol.*, 197, 286–292, doi:10.1016/S0022-1694(96)03276-3, 1997.
- Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, *Water Resour. Res.*, 35, 2739–2750, doi:10.1029/1999wr900099, 1999.
- Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, *J. Hydrol.*, 249, 2–9, doi:10.1016/S0022-1694(01)00420-6, 2001.
- Krzysztofowicz, R. and Kelly, K. S.: Hydrologic uncertainty processor for probabilistic river stage forecasting, *Water Resour. Res.*, 36, 3265–3277, doi:10.1029/2000WR900108, 2000.
- Kuczera, G.: Improved Parameter Inference in Catchment Models. 1. Evaluating Parameter Uncertainty, *Water Resour. Res.*, 19, 1151–1162, doi:10.1029/WR019i005p01151, 1983.
- Kuczera, G., Kavetski, D., Franks, S., and Thyer, M.: Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, 331, 161–177, doi:10.1016/j.jhydrol.2006.05.010, 2006.
- Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.
- Li, M., Wang, Q. J., and Bennett, J. C.: Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, *Water Resour. Res.*, 49, 5913–5929, doi:10.1002/wrcr.20445, 2013.
- Li, M., Wang, Q. J., Bennett, J. C., and Robertson, D. E.: A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts, *Hydrol. Earth Syst. Sci.*, 19, 1–15, doi:10.5194/hess-19-1-2015, 2015.
- Marshall, L., Sharma, A., and Nott, D.: Modeling the catchment via mixtures: Issues of model specification and validation, *Water Resour. Res.*, 42, W11409, doi:10.1029/2005WR004613, 2006.
- Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, 40, W01106, doi:10.1029/2003wr002540, 2004.
- Moradkhani, H., Hsu, K. L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, 41, W05012, doi:10.1029/2004wr003604, 2005.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.
- Nelder, J. A. and Mead, R.: A Simplex Method for Function Minimization, *The Computer Journal*, 7, 308–313, doi:10.1093/comjnl/7.4.308, 1965.
- NOAA National Weather Service: MOPEX (Model Parameter Estimation Experiment), available at: http://www.nws.noaa.gov/ohd/mopex/mo_datasets.htm, last access: August 2016.
- Peng, Z. L., Wang, Q. J., Bennett, J. C., Schepen, A., Pappenberger, F., Pokhrel, P., and Wang, Z. R.: Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China, *J. Geophys. Res.-Atmos.*, 119, 7116–7135, doi:10.1002/2013JD021162, 2014.
- Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.
- Pianosi, F. and Raso, L.: Dynamic modeling of predictive uncertainty by regression on absolute errors, *Water Resour. Res.*, 48, W03516, doi:10.1029/2011wr010603, 2012.
- Reggiani, P., Renner, M., Weerts, A. H., and van Gelder, P. H. A. J. M.: Uncertainty assessment via Bayesian revision of ensemble streamflow predictions in the operational river Rhine forecasting system, *Water Resour. Res.*, 45, W02428, doi:10.1029/2007wr006758, 2009.
- Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, 46, W05521, doi:10.1029/2009wr008328, 2010.
- Robertson, D. E., Shrestha, D. L., and Wang, Q. J.: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting, *Hydrol. Earth Syst. Sci.*, 17, 3587–3603, doi:10.5194/hess-17-3587-2013, 2013.
- Schaeffli, B., Talamba, D. B., and Musy, A.: Quantifying hydrological modeling errors through a mixture of normal distributions, *J. Hydrol.*, 332, 303–315, doi:10.1016/j.jhydrol.2006.07.005, 2007.
- Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, 46, W10531, doi:10.1029/2009wr008933, 2010.
- Shrestha, D. L., Robertson, D. E., Bennett, J. C., and Wang, Q. J.: Improving Precipitation Forecasts by Generating Ensembles through Postprocessing, *Mon. Weather Rev.*, 143, 3642–3663, doi:10.1175/MWR-D-14-00329.1, 2015.
- Shrestha, D. L., Robertson, D. E., Wang, Q. J., Pagano, T. C., and Hapuarachchi, H. A. P.: Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose, *Hydrol. Earth Syst. Sci.*, 17, 1913–1931, doi:10.5194/hess-17-1913-2013, 2013.
- Smith, T., Sharma, A., Marshall, L., Mehrotra, R., and Sisson, S.: Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resour. Res.*, 46, W12551, doi:10.1029/2010wr009514, 2010.
- Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, *Water Resour. Res.*, 45, W00B11, doi:10.1029/2008wr006839, 2009.
- Sorooshian, S. and Dracup, J. A.: Stochastic parameter-estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, 16, 430–442, doi:10.1029/WR016i002p00430, 1980.
- Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, *Hydrol. Earth Syst. Sci.*, 13, 125–140, doi:10.5194/hess-13-125-2009, 2009.
- Thyer, M., Kuczera, G., and Wang, Q. J.: Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation, *J. Hydrol.*, 265, 246–257, doi:10.1016/S0022-1694(02)00113-0, 2002.
- Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, *International Journal of River Basin Management*, 6, 123–137, doi:10.1080/15715124.2008.9635342, 2008.
- Vaze, J., Perraud, J. M., Teng, J., Chiew, F. H. S., Wang, B., and Yang, Z.: Catchment Water Yield Estimation Tools (CWYET), in: *Proceedings of the 34th World Congress of the International*

- Association for Hydro-Environment Research and Engineering: 33rd Hydrology and Water Resources Symposium and 10th Conference on Hydraulics in Water Engineering, edited by: Valentine, E., Apelt, C., Ball, J., Chanson, H., and Sargison, J., Engineers Australia, Brisbane, 1554–1561, 2011.
- Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, 41, W01017, doi:10.1029/2004wr003059, 2005.
- Wang, Q. J. and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, *Water Resour. Res.*, 47, W02546, doi:10.1029/2010WR009333, 2011.
- Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resour. Res.*, 45, W05407, doi:10.1029/2008WR007355, 2009.
- Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, *Water Resour. Res.*, 48, W05514, doi:10.1029/2011WR010973, 2012.
- Wang, Q. J., Bennett, J. C., Schepen, A., Robertson, D. E., Song, Y., and Li, M.: FoGSS – A model for generating forecast guided stochastic scenarios of monthly streamflows out to 12 months, CSIRO Water for a Healthy Country Flagship, Highett, Australia, Report, 34 pp., 2014.
- Xiong, L. H. and O'Connor, K. M.: Comparison of four updating models for real-time river flow forecasting, *Hydrolog. Sci. J.*, 47, 621–639, doi:10.1080/02626660209492964, 2002.
- Yang, J., Reichert, P., Abbaspour, K. C., and Yang, H.: Hydrological modelling of the chaohe basin in china: Statistical model formulation and Bayesian inference, *J. Hydrol.*, 340, 167–182, doi:10.1016/j.jhydrol.2007.04.006, 2007.
- Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q., and Zhao, J.: Quantifying predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model, *J. Hydrol.*, 528, 329–340, doi:10.1016/j.jhydrol.2015.06.043, 2015.