A strategy to overcome adverse effects of autoregressive updating of streamflow forecasts

M. Li[1], Q. J. Wang[2], J. C. Bennett[2] and D. E. Robertson[2]

[1] CSIRO Computational Informatics, Floreat, Western Australia, Australia

[2] CSIRO Land and Water, Highett, Victoria, Australia

Correspondence to: M. Li (Ming.Li@csiro.au)

Abstract

For streamflow forecasting applications, rainfall-runoff hydrological models are often augmented with updating procedures that correct streamflow forecasts based on the latest available observations of streamflow and their departures from model simulations. The most popular approach uses autoregressive (AR) models that exploit the "memory" in hydrological model simulation errors. AR models may be applied to raw errors directly or to normalised errors. In this study, we demonstrate that AR models applied in either way can sometimes cause over-correction of forecasts. In using an AR model applied to raw errors, the over-correction usually occurs when streamflow is rapidly receding. In applying an AR model to normalised errors, the over-correction usually occurs when streamflow is rapidly rising. Furthermore, when parameters of a hydrological model and an AR model are estimated jointly, the AR model applied to normalised errors sometimes degrades the stand-alone performance of the base hydrological model. This is not desirable for forecasting applications, as forecasts should rely as much as possible on the base hydrological model, and updating should be applied only to correct minor errors. To overcome the adverse effects of the conventional AR models, a restricted AR model applied to normalised errors is introduced. The new model is evaluated on a number of catchments and is shown to reduce over-correction and to improve the performance of the base hydrological model considerably.

1. Introduction

Rainfall-runoff models are widely used to generate streamflow forecasts, which provide essential information for flood warning and water resources management. For streamflow forecasting, rainfall-runoff models are often augmented by updating procedures that correct streamflow forecasts based on the latest available observations of streamflow and their departures from model simulations. Model errors reflect limitations of the hydrological models in reproducing physical processes as well as inaccuracies in data used to force and evaluate the models.

The most popular updating approach uses autoregressive (AR) models, which exploit the "memory" - more precisely the autocorrelation structure - of errors in hydrological simulations (Morawietz et al., 2011). Essentially, AR updating uses a linear function of the known errors at previous time steps to anticipate errors in a forecast period. Forecasts are then updated according to these anticipated errors. AR updating is conceptually simple and yet generally leads to significantly improved forecasts (World Meteorological Organization, 1992). AR updating has been shown to provide equivalent performance to more sophisticated non-linear and nonparametric updating procedures (Xiong and O'Connor, 2002).

In rainfall-runoff modelling, model errors are generally heteroscedastic (i.e., they have heterogeneous variance over time) (Xu, 2001;Kavetski et al., 2003;Pianosi and Raso, 2012) and non-Gaussian (Bates and Campbell, 2001;Schaefli et al., 2007;Shrestha and Solomatine, 2008). In many applications (Seo et al., 2006;Bates and Campbell, 2001;Salamon and Feyen, 2010;Morawietz et al., 2011), AR models are applied to normalised errors that are considered homoscedastic and Gaussian. Normalisation is often achieved through variable transformation by using, for example, the Box-Cox transformation (Thyer et al., 2002;Bates and Campbell, 2001;Engeland et al., 2010) or, more recently, the log-sinh transformation (Wang et al., 2012;Del Giudice et al., 2013). In other applications (Schoups and Vrugt, 2010;Schaefli et al., 2007), AR models are applied directly to raw errors, but residual errors of the AR models may be explicitly specified as heteroscedastic and non-Gaussian.

There is no agreement on whether it is better to apply an AR model to normalised or raw errors. Recent work by Evin et al. (2013) found that an AR model applied to raw errors may lead to poor performance with exaggerated uncertainty. They demonstrated that such instability can be mitigated by applying an AR model to standardised errors (raw errors divided by standard deviations). Here, standardisation has a similar effect to normalisation in that it homogenises the variance of the errors (but does not consider the non-Gaussian distribution of errors). Conversely, Schaefli et al. (2007) pointed out that when an AR model is jointly estimated with a hydrological model, there is a clear advantage in applying an AR model to raw errors rather than normalised (or standardised) errors. Schaefli et al. (2007) found that using raw errors leads to more reliable parameter inference and uncertainty estimation, because the mean error is close to zero and therefore the simulations are free of systematic bias. The same is not necessarily true when applying an AR model to normalised errors.

In this study, we evaluate AR models applied to both raw and normalised errors on four Australian catchments and three United States (US) catchments. We show that when estimated jointly with a hydrological model, the AR model applied to normalised errors sometimes degrades the stand-alone performance of the base hydrological model. We also identify that both of these conventional AR models can sometimes cause over-correction of forecasts. We introduce a restricted AR model applied to normalised errors and demonstrate its effectiveness in overcoming the adverse effects of the conventional AR models.

2. Autoregressive error models

2.1 Formulations

A hydrological model is a function of forcing variables (precipitation and potential evapotranspiration), initial catchment state, S_0, and a set of hydrological model parameters, θ_H. We denote the observed streamflow and model simulated streamflow at time t by Q_t and $\tilde{Q}_t$, respectively. An error model is used to describe the difference between Q_t and $\tilde{Q}_t$. The log-sinh transformation defined by Wang et al. (2012)

$$f(x) = b^{-1}\log\{\sinh(a+bx)\} \tag{1}$$

is applied to stabilise variance and normalise data.

In this study, we firstly examine two first-order AR error models:

(1) An AR error model applied to normalised errors (referred to as *AR-Norm*) defined by:

$$Z_t = \tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right) + \varepsilon_t \,, \tag{2}$$

where Z_t and $\tilde{Z}_t$ are the log-sinh transformed variables of Q_t and $\tilde{Q}_t$;

(2) An AR error model applied to raw errors (referred to as *AR-Raw*) defined by

$$Z_t = f\left\{\tilde{Q}_t + \rho\left(Q_{t-1} - \tilde{Q}_{t-1}\right)\right\} + \varepsilon_t \,. \tag{3}$$

For both models, ρ is the lag-1 autoregression parameter, and ε_t is an identically and independently distributed Gaussian deviate with a mean of zero and a constant standard deviation σ.

Both the AR-Norm and AR-Raw models represent the lag-one autocorrelation by an AR process and both employ the log-sinh transformation. However, the way the log-sinh transformation is applied differs between the two models. The AR-Norm model first applies the log-sinh transformation to the observed and model simulated streamflow, and then assumes that the error in the transformed space follows an AR(1) process. In contrast, the AR-Raw model essentially assumes that the error in the original space follows an AR(1) process and only applies the log-sinh transformation to fit the asymmetric and non-Gaussian error distribution.

The median of the updated streamflow forecast (referred to as *updated streamflow*) for the AR-Norm and AR-Raw models (see Appendix A for proof), denoted by $\tilde{Q}_t^*$, are respectively

$$\tilde{Q}_t^* = f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right)\right\}, \tag{4}$$

and

$$\tilde{Q}_t^* = \tilde{Q}_t + \rho\left(Q_{t-1} - \tilde{Q}_{t-1}\right), \tag{5}$$

where $f^{-1}(x)$ is the inverse of log-sinh transformation (or back-transformation). The magnitude of the error update by the AR-Raw model, $\tilde{Q}_t^* - \tilde{Q}_t$, is dependent only on the difference between Q_{t-1} and $\tilde{Q}_{t-1}$. In contrast, the magnitude of the error update by the AR-Norm model is dependent not only on the difference between Q_{t-1} and $\tilde{Q}_{t-1}$, but also on $\tilde{Q}_t$. Put differently, the AR-Norm model uses errors calculated in the transformed domain, and this means that the error in the original domain can be amplified (or reduced) by the back-transformation (Equation (4)). The AR-Raw model uses errors calculated in the original domain and no back-transformation is used in calculating $\tilde{Q}_t^*$ (Equation (5)), meaning that the error in the original domain cannot be amplified (or reduced). In Appendix B, we show that the AR-Norm model gives greater error updates for larger values of $\tilde{Q}_t$.

We will demonstrate in Section 4 that the AR-Norm and AR-Raw models can sometimes cause over-correction of forecasts. Motivated to overcome the potential for over-correction, we introduce a modification of the AR-Norm model, called the restricted AR-Norm model (referred to as *RAR-Norm*). A condition $\left|\tilde{Q}_t^* - \tilde{Q}_t\right| \le \left|Q_{t-1} - \tilde{Q}_{t-1}\right|$ is used to limit the correction amount to not exceeding the error in the last time step in absolute value. The updated streamflow is given by

$$\tilde{Q}_t^* = \begin{cases} f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right)\right\} & \text{if } D_t \le \left|Q_{t-1} - \tilde{Q}_{t-1}\right| \\ \tilde{Q}_t + (Q_{t-1} - \tilde{Q}_{t-1}) & \text{otherwise.} \end{cases} \tag{6}$$

where

$$D_t = \left| f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right)\right\} - \tilde{Q}_t \right|. \tag{7}$$

The full RAR-Norm model in the transformed space is given by

$$Z_t = \begin{cases} \tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right) + \varepsilon_t & \text{if } D_t \le \left|Q_{t-1} - \tilde{Q}_{t-1}\right| \\ f\left(\tilde{Q}_t + Q_{t-1} - \tilde{Q}_{t-1}\right) + \varepsilon_t & \text{otherwise.} \end{cases} \tag{8}$$

2.2 Estimation

The AR-Norm, AR-Raw and RAR-Norm models are each calibrated jointly with the hydrological model. The method of maximum likelihood is used to estimate the error model parameters θ_E and the hydrological model parameters θ_H . Using a similar derivation as given by Li et al. (2013), the likelihood functions can be written as

(a) for AR-Norm

$$L\left(\theta_E,\theta_H\right)=\prod_t P(Q_t \mid \tilde{Q}_t,\tilde{Q}_{t-1};\theta_E,\theta_H)=\prod_t J_{Z_t\to Q_t}\phi\left(Z_t \mid \tilde{Z}_t+\rho\left(Z_{t-1}-\tilde{Z}_{t-1}\right),\sigma^2\right), \quad (9)$$

(b) for AR-Raw

$$L\left(\theta_E,\theta_H\right)=\prod_t P(Q_t \mid \tilde{Q}_t,\tilde{Q}_{t-1};\theta_E,\theta_H)=\prod_t J_{Z_t\to Q_t}\phi\left(Z_t \mid f\left\{\tilde{Q}_t+\rho\left(Q_{t-1}-\tilde{Q}_{t-1}\right)\right\},\sigma^2\right) \quad , \quad (10)$$

(c) for RAR-Norm

$$\begin{aligned}L\left(\theta_E,\theta_H\right)=\prod_t P(Q_t \mid \tilde{Q}_t,\tilde{Q}_{t-1};\theta_E,\theta_H)=&\prod_{t:D_t\le\left|Q_{t-1}-\tilde{Q}_{t-1}\right|} J_{Z_t\to Q_t}\phi\left(Z_t \mid \tilde{Z}_t+\rho\left(Z_{t-1}-\tilde{Z}_{t-1}\right),\sigma^2\right)\\ &+\prod_{t:D_t>\left|Q_{t-1}-\tilde{Q}_{t-1}\right|} J_{Z_t\to Q_t}\phi\left(Z_t \mid f\left\{\tilde{Q}_t+\rho\left(Q_{t-1}-\tilde{Q}_{t-1}\right)\right\},\sigma^2\right),\end{aligned} \quad (11)$$

where $J_{Z_t\to Q_t}=\left\{\tanh(a+bQ_t)\right\}^{-1}$ is the Jacobian determinant of the log-sinh transformation and $\phi\left(x \mid \mu,\sigma^2\right)$ is the probability density function of a Gaussian random variable x with mean μ and standard deviation σ. The probability density function is replaced by the cumulative probability function when evaluating events of zero flow occurrences (Wang and Robertson, 2011;Li et al., 2013). The Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1994) is used to minimize the negative log likelihood.

3. Data

We use daily data from four Australian catchments and three catchments from the United States (US; Figure 1, Table 1). Australian streamflow data are taken from the Catchment Water Yield Estimation Tool (CWYET) dataset (Vaze et al., 2011). Australian rainfall and potential evaporation data are derived from the Australian Water Availability Project (AWAP) dataset (Jones et al., 2009). All data for the US catchments come from the Model Intercomparison Experiment (MOPEX) dataset (Duan et al., 2006). The selected US catchments are amongst the 12 catchments used by Evin et al. (2014) to compare joint and postprocessor approaches to estimate hydrological uncertainty, and allows us to compare results with that study (the other catchments used by Evin et al. (2014) are influenced by snowmelt, which is not considered in the hydrological model used in this study). The Abercrombie River and the Guadalupe River intermittently experience periods of very low (to zero) flow, while the other rivers flow perennially (Table 1). Such dry catchments are challenging for hydrological simulations and error modelling. All catchments have high-quality streamflow records with very few missing data.

We forecast daily streamflow with the GR4J rainfall-runoff model (Perrin et al., 2003) . We apply updating procedures to correct these forecasts. All results presented in this paper are based on this cross-validation instead of calibration in order to ensure the results can be generalised to independent data. We use different cross-validation schemes for the Australian and US catchments, because of the shorter streamflow records available for the Australian catchments:

i. For the Australian catchments we use data from 1992 to 2005 (14 years) for these catchments. We then generate 14-fold cross-validated streamflow forecasts. The data from 1990-1991 are only used to warm up the GR4J model. For a given year, we leave out the data from that year and the following year when estimating the parameters of GR4J and error models. For example, if we wish to forecast streamflows at any point in 1999, we leave out data from 1999 and 2000 when we estimate parameters. The removal of data from the following year (2000) is designed to minimise the impact of hydrological memory on model parameter estimation. We then generate streamflow forecasts in that year (1999) with model parameters estimated from the remaining data.
ii. For the US catchments we follow the split-sampling validation scheme suggested by Evin et al. (2014) to make our results comparable to that study: (1) an 8-year calibration (09/09/1973- 26/11/1981) (i.e. 3000 days) with an 8-year warm-up period and (2) a 17-year validation (27/11/1981-01/05/1998) (i.e. 6000 days) with an 8-year warm-up period.

To demonstrate the problems of over-correction of errors in updating and poor stand-alone performance of the base hydrological model, we consider only streamflow forecasts for one time step ahead. We will consider longer lead times in future work. Forecasts are generated using observed rainfall (i.e., a ‘perfect’ rainfall forecast) as input. In streamflow forecasting, forecasts may be generated from rainfall information that comes from a different source (e.g., a numerical weather prediction model). Our study is aimed at streamflow forecasting applications, so we preserve the distinction between observed and forecast forcings by referring to streamflows modelled with observed rainfall as *simulations* and those modelled with forecast rainfall as *forecasts*. As the forecast rainfall we use is observed rainfall, the terms *forecast* and *simulation* are interchangeable.

4. Results

4.1 Over-correction of forecasts as the hydrograph rises

The first adverse effect of the conventional AR models is over-correction of errors in updating as streamflows are rising. By over-correction, we mean that the AR model updates the hydrological model simulations too much. Over-correction is difficult to define precisely, however we will demonstrate the concept with two examples in the Mitta Mitta catchment: the first example illustrates over-correction by the AR-Norm model, and the second example illustrates over-correction by the AR-Raw model.

To illustrate the problem of over-correction caused by the AR-Norm model, Figure 2 presents a 1-week time series for the Mitta Mitta catchment, showing streamflow forecasts with GR4J before error updating (referred to as streamflows forecast with the *base hydrological model*) and after error updating. Figure 2 shows that the base hydrological models consistently under-estimate the streamflow from 23/09/2000 to 25/09/2000, and the corresponding updating procedures successfully identify the need to compensate for this under-estimation. For the AR-Norm model, however, the correction amount for 26/09/2000 is unreasonably large. Because the forecast streamflow on 26/09/2000 is much higher than that of the previous day, the correction is greatly amplified by the back-transformation, leading to the over-correction. In contrast, the AR-Raw model works better in this situation because the magnitude of the error update never exceeds the simulation error on the previous day regardless of whether the forecast streamflow is high or low. The RAR-Norm model behaves similarly to the AR-Raw model for correcting the peak on 26/09/2000 and avoids the over-correction made by the AR-Norm model.

Figure 3 shows instances of possible over-correction by the AR-Norm model, identified by the condition $D_t > \left| Q_{t-1} - \tilde{Q}_{t-1} \right|$. Figure 3 shows that about 10-25% of the AR-Norm updated forecasts have an error update that is larger than the forecast error on the previous day and therefore are susceptible to over-correction. The frequency of these instances varies somewhat from catchment to catchment. The RAR-Norm model identifies 10-30% of the forecasts as possible instances of problematic updating, and the AR-Norm model identifies a similar number of instances (slightly fewer – they are not identical because the parameters for each model are inferred independently).

Figure 4 presents a time-series for the Orara catchment that shows the instances susceptible to over-correction for the AR-Norm model. These instances all occur when the streamflow rises. The RAR-Norm model effectively rectifies the problem of over-correction caused by the AR-Norm model. We note that there is nothing that forces the instances susceptible to over-correction identified by the AR-Norm model to be the same as those identified by the RAR-Norm models because the two models are calibrated independently (and therefore base hydrological model simulations may be different). However, the restriction defined in the RAR-Norm model is largely applied to the instances where the AR-Norm model is susceptible to over-correction.

4.2 Over-correction of forecasts as the hydrograph recedes

The second adverse effect of conventional AR models is over-correction of forecasts as streamflows reced. Anexample is presented in Figure 5 where the AR-Raw model causes over-correction. Here, the base hydrological model over-estimates the receding hydrograph on 05/10/1993. The magnitude of the error update given by the AR-Raw model cannot adjust according to the value of the forecast. As a result, the AR-Raw model updates the forecast on 06/10/1993 by a large amount, resulting in serious under-estimation (the forecast is for near zero streamflow), and an artificial distortion of the hydrograph. (We note that we have seen this problem become much worse in unpublished experiments of forecasts made for several time-steps into the future, sometimes resulting in forecasts of zero flows during large floods.) In contrast, the AR-Norm model performs better in this example, giving a smaller magnitude of error update by recognising that the hydrograph is moving downward. It is generally true that in applying the AR-Raw model, over-correction may occur when the streamflow is receding. The RAR-Norm model produces updated streamflow similar to the AR-Norm model when the hydrograph recedes rapidly and avoids the over-correction by the AR-Raw model on 06/10/1993.

Figure 6 provides more examples of the over-correction caused by the AR-Raw model from a longer time-series plot for the Abercrombie catchment. There are three clear instances of over-correction, all occurring on the time step immediately after large peaks in observed streamflows. The RAR-Norm works better than the AR-Raw model to avoid the three instances of over-correction for the Abercrombie catchment. Overall, the RAR-Norm model takes a conservative position when streamflow changes rapidly, either rising or falling. When streamflow changes rapidly, it is difficult to anticipate the magnitude of forecast error. Accordingly the conventional AR models are prone to over-correction in such instances.

4.3 Poor stand-alone performance of the base hydrological model

The third adverse effect with conventional AR error models is the stand-alone performance of the base hydrological model (GR4J). As noted above, the parameters of the base hydrological model are estimated jointly with each error model. For streamflow forecasting, we expect to obtain a reasonably accurate forecast from the base hydrological model followed by an updating procedure as an auxiliary means to improve the forecast accuracy. At lead times of many time-steps (e.g., streamflow forecasts generated from medium-range rainfall forecasts) the magnitude of AR error updates becomes rapidly smaller (tending to zero), and thus the performance of the base hydrological model is crucial for realistic forecasts at longer lead times. While we investigate only forecasts at a lead time of one time step in this study, we aim to develop methods that can be applied to forecasts at longer lead times. Further, if the base hydrological model does not replicate important catchment processes realistically, the performance of the hydrological model outside the calibration period may be less robust.

Figure 7 presents the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) calculated from the base hydrological model and the error models. When the AR-Norm model is used, the forecasts from the base hydrological model are very poor for the Orara catchment (NSE<0). The scatter plot in Figure 8 shows a serious over-estimation of the streamflow simulation for the Orara. When the AR-Norm model is used, the base hydrological model greatly over-estimates discharge and the AR-Norm model then attempts to correct this systematic over-estimation. This is also shown in Figure 4 where the base hydrological model has a strong tendency to over-estimate streamflows for a range of streamflow magnitudes. The base hydrological model with the AR-Norm model also performs poorly for the Abercrombie catchment (Figure 7). In this case, the base hydrological model tends to under-estimate streamflows (results not shown). For the other three catchments, however, the base hydrological model with the AR-Norm model performs reasonably well.

In general, the AR-Raw base hydrological model performs as well or better than the AR-Norm base hydrological model. The AR-Raw base hydrological model is notably better than the AR-Norm base hydrological model in the Abercrombie and Orara catchments (Figure 7). This suggests that more robust performance can be expected of base hydrological models when AR models are applied to raw errors.

The RAR-Norm model generally improves the performance of the AR-Norm base hydrological model to a similar performance level of the AR-Raw base hydrological model (Figure 7). The improvement over the AR-Norm base hydrological model is especially evident for the Orara (Figures 4 and 7) and Abercrombie catchments (Figures 7).

We note that for the AR-Norm models, the updated forecasts are not always better than forecasts generated by the base hydrological models. For the Tarwin and Guadalupe catchments, AR-Norm forecasts are not as good as the forecasts generated by the AR-Norm base hydrological model. This points to a tendency to overfit the parameters to the calibration period, resulting in the error model undermining the performance of the base hydrological model under cross-validation. Such a lack of robustness is highly undesirable in forecasting applications, where the hydrological models should be able to operate in conditions that differ from those experienced during calibration. Note that this problem also occurs in the RAR-Norm model (Guadalupe) and in the AR-Raw model (Abercrombie, Guadalupe) but to a much smaller degree.

In general, the updated forecasts from the RAR-Norm model show similar or better forecast accuracy, as measured by NSE, than both the AR-Raw model and the AR-Norm model (Figure 7). We note that the Orara catchment is an exception: here the AR-Raw model shows slightly better performance than RAR-Norm model. Conversely, the RAR-Norm model shows notably better performance than both the AR-Norm and AR-Raw models in the Abercrombie and Guadalupe catchments. This suggests the RAR-Norm model may work better in intermittently flowing catchments, although further testing is required to establish that this is true for a greater range of catchments.

4.4 Further analyses

We further evaluate the NSE of the three different error models calibrated when streamflows are receding (i.e. $\tilde{Q}_t \le \tilde{Q}_{t-1}$) and rising (i.e. $\tilde{Q}_t > \tilde{Q}_{t-1}$) (Table 2). For the receding streamflows (constituting 70-85% of streamflows), the AR-Raw model leads to the overall worst forecast accuracy because of the over-correction explained in Section 4.1. This is especially evident for the Abercrombie catchment (and, to a lesser degree, the Guadalupe catchment). The RAR-Norm model significantly outperforms the other two models for the Abercrombie catchment and shares similar forecast accuracy to the (strongly performing) AR-Norm model for the other catchments. When streamflows are rising (which also includes streamflow peaks), the AR-Norm model can cause over-correction and leads to the least accurate forecasts (in terms of NSE), and the RAR-Norm model behaves similarly to the AR-Raw model, which consistently provides the most accurate forecasts. (The only exception is the Guadalupe River, where the AR-Raw model clearly outperforms the RAR-Norm model when streamflows are rising. This is somewhat compensated for by the markedly better performance the RAR-Norm model offers over the AR-Raw model when streamflows are receding for this catchment, leading to better forecasts overall (Figure 7).) We conclude that the AR-Norm model generally tends to perform least well when streamflows recede, and that the AR-Raw model tends to perform least well when streamflows rise. We also conclude that the RAR-Norm model tends to combine the best elements of the AR-Norm and AR-Raw models, leading to the best overall performance.

We have shown that over-corrections can lead to inaccurate deterministic forecasts, and we now discuss the consequences for the probabilistic predictions given by each of the error models. We assess probabilistic forecast skill with skill scores derived from two probabilistic verification measures: the Continuous Rank Probability Score (CRPS) and the Root Mean Square Error in Probability (RMSEP) (denoted by CRPS_SS and RMSEP_SS, respectively) (Wang and Robertson, 2011). Both skill scores are calculated with respect to a reference forecast. The reference forecast is generated by resampling historical streamflows: for a forecast issued for a given month/year (e.g. February 1999), we randomly draw a sample of 1000 daily streamflows that occurred in that month (e.g. February) from other years with replacement (e.g. years other than 1999). Table 3 compares these two skill scores calculated for the all catchments. The RAR-Norm model performs best across the range of skill scores and catchments, attaining the highest CRPS_SS in 4 of the 7 catchments and the highest RMSEP_SS in 4 of 7 catchments. Even where RAR-Norm was not the best performed model, it performs very similarly to the best performing model in all cases. Interestingly, the AR-Raw model tends to outperform the AR-Norm model in CRPS_SS while the reverse is true for RMSEP_SS. The CRPS tests how appropriate the spread of uncertainty is for each probabilistic forecast, while RMSEP puts little weight on this. The results suggest that while the median forecasts of AR-Norm tends to be slightly more accurate than those of the AR-Raw model, the forecast uncertainty is represented slightly better by the AR-Raw model.

To better understand how reliably the forecast uncertainty is quantified by each model, we produce Probability Integral Transform (PIT) uniform probability plots (Wang and Robertson, 2011) in Figure 9. There are two main points to draw from these plots. First, the curves are very similar for all error models (a partial exception is the San Marcos catchment, where the AR-Raw model is slightly closer to the one-to-one line than the other models). This demonstrates that in general the models produce similarly reliable uncertainty distributions. Second, all models show an inverted S-shaped curve, which is characteristic of the forecasts with uncertainty ranges that are too wide. This underconfidence is a result of using a Gaussian distribution to characterise the error. The Gaussian distribution is not flexible enough to represent the high degree of kurtosis in the distribution of the residuals after error updating (partly because the errors become very small after updating). We are presently experimenting with other distributions in order to address this issue, and will seek to publish this work in future. For the purposes of the present study, we conclude that the three error models are similarly reliable.

5. Discussion and conclusions

For streamflow forecasting, rainfall-runoff models are often augmented with an updating procedure that corrects the forecast using information from recent simulation errors. The most popular updating approach uses autoregressive (AR) models that exploit the "memory" in model errors. AR models may be applied to raw errors directly or to normalised errors.

We demonstrate three adverse effects of AR error updating procedures on seven catchments. The first adverse effect is possible over-correction on the rising limb of the hydrograph. The AR-Norm model can exhibit the tendency to over-correct the peaks or on the rise of a hydrograph, because error updating can be (overly) amplified by the back-transformation. The second adverse effect is the tendency to over-correct receding hydrographs. This tendency is most prevalent in the AR-Raw model, which can fail to recognise that a large error update may not be appropriate for small streamflows.

The third adverse effect is that the stand-alone performance of base hydrological models can be poor when the parameters of rainfall-runoff and error models are jointly estimated with the AR parameters. We show that poor base hydrological model performance is particularly prevalent in the AR-Norm model. The poor performance appears to occur in catchments with highly skewed streamflow observations (the intermittent Abercrombie River, and the Orara River, a catchment in a subtropical climate). For example, in the Orara River, the base hydrological model tends to greatly over-estimate streamflows, and then relies on the error updating to correct the over-estimates. This is not desirable in real-time forecasting applications for two major reasons. First, modern streamflow forecasting systems often extend forecast lead-times with rainfall forecast information (Bennett et al., 2014). The magnitude of AR updating decays with lead times, and forecasts at longer lead times rely heavily on the performance of the base hydrological model. Second, hydrological models are designed to simulate various components of natural systems, such as baseflow processes or overland flow. In theory, simulating these processes correctly will allow the model to perform well for climate conditions that may substantially differ from those experienced during the parameter estimation period. If the hydrological model parameters do not reflect the natural processes for a given catchment, the hydrological model may be much less robust outside the parameter estimation period.

We note that the poor performance of the hydrological model may be specific to the GR4J model, and many not occur in other hydrological models. Evin et al. (2014) estimated hydrological model and error model parameters jointly using GR4J and another hydrological model, HBV, for the three US catchments tested here. While they did not assess the performance of the base hydrological models, they found that HBV tended to perform more robustly when combined with different error models. It is possible that we may have achieved more stable base model performance had we used HBV or another hydrological model. We note, however, that our conclusions can probably be generalised to other hydrological models that do not offer robust base model performance under joint parameter estimation (e.g. GR4J). Because the RAR-Norm model essentially limits the range of updating that can be applied through the AR-Norm model, it will tend to rely more heavily on the base hydrological model, and therefore will tend to favour parameter sets that encourage good stand-alone performance of the base model. For those hydrological models that already produce robust base model performance under joint parameter estimation (perhaps HBV), RAR-Norm is unlikely to undermine this performance for the same reasons. We see some evidence of this in our experiments with GR4J: when the performance of the base hydrological model is already strong relative to the updated forecasts for the AR-Norm and AR-Raw models (e.g. the Tarwin, Mitta Mitta, or Guadalupe catchments), the RAR-Norm model base hydrological model also performs strongly.

The tendency of the AR-Norm model to over-correct rising streamflows is probably generic. In particular, transformations other than the log-sinh transformation may still lead to over-correction at the peak of hydrograph. The proof in Appendix A shows that if a transformation satisfies some conditions (first derivate is positive and second derivate is negative), it will tend to correct more for higher forecast streamflows and can cause the problem of over-correction. The conditions given by Appendix A are generally true for many other transformations used for data normalisation and variance stabilisation in hydrological applications, such as logarithm transformation and Box-Cox transformation with the power parameter less than 1.

We use joint parameter inference to calibrate hydrological model and error model parameters, in order to address the true nature of underlying model errors. Inferring parameters of the error model and the base hydrological model independently – i.e., first inferring parameters of the base hydrological model, holding these constant and then inferring the error model parameters - relies on simplified and often invalid error assumptions (it assumes independent, homoscedastic and Gaussian errors), but nonetheless could be a pragmatic alternative to the joint parameter inference to reduce computational demands. The over-correction of conventional AR models is independent of the parameter inference, whether the error and base hydrological model parameters are inferred jointly or independently.

In order to mitigate the adverse effects of conventional AR updating procedures, we introduce a new updating procedure called the RAR-Norm model. The RAR-Norm model is a modification of the AR-Norm: in most instances it operates as the AR-Norm model, but in instances of possible over-correction it relies on the error in untransformed streamflows at the previous time step. That is, RAR-Norm is essentially a more conservative error model than AR-Norm: in situations where streamflows change rapidly, it opts to update with whichever error (transformed or untransformed) is smaller. This forces greater reliance on the base hydrological model to simulate streamflows accurately, leading to more robust performance in the base hydrological model. The RAR-Norm model clearly outperforms the AR-Norm model in both the updated and base model forecasts, as well as ameliorating the problem of over-correcting rising streamflows. The RAR-Norm model's advantage over the AR-Raw model is less clear: both the base hydrological model and the updated forecasts produced by the AR-Raw model perform similarly to (or sometimes slightly better than) the RAR-Norm model. However, the RAR-Norm model clearly addresses the problem of over-correcting receding streamflows that occurs in the AR-Raw model. As we show, this type of over-correction can seriously distort event hydrographs, and cause forecasts of near zero streamflows when reasonably substantial streamflows are observed. While these instances are not very common, the failure in the forecast is a serious one. As we note earlier, the over-correction of receding streamflows is likely to be exacerbated when producing forecasts at lead times of more than one time step. Accordingly, we contend that the RAR-Norm model is preferable to both AR-Norm and AR-Raw models for streamflow forecasting applications.

Appendix A

For simplicity we only show the case of the AR-Norm model and analogues arguments can be used to prove the cases of the AR-Raw and RAR-Norm models. The streamflow ensemble forecast Q_t given by the AR-Norm model defined by (1) can be written as

$$Q_t = \max\left[f^{-1}\left\{ \tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1} \right) + \varepsilon_t \right\}, 0 \right]. \tag{A1}$$

where negative values after the back-transformation are assigned zero values. Because we assume that ε_t is a standard normal random variable, In order to show $\tilde{Q}_t^*$ is the median of Q_t, we just need to show $P\left(Q_t \le \tilde{Q}_t^*\right) = 0.5$, which can be proved as follows:

$$\begin{aligned} P\left(Q_t \le \tilde{Q}_t^*\right) &= P\left(\max\left[f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right) + \varepsilon_t\right\}, 0\right] \le \tilde{Q}_t^*\right) \\ &= P\left(f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right) + \varepsilon_t\right\} \le \tilde{Q}_t^* \text{ and } 0 \le \tilde{Q}_t^*\right) \end{aligned} \tag{A2}$$

Because $\tilde{Q}_t^*$ always has a non-negative value, we have

$$\begin{aligned} P\left(Q_t \le \tilde{Q}_t^*\right) &= P\left(f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right) + \varepsilon_t\right\} \le f^{-1}\left\{\tilde{Z}_t + \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right)\right\}\right) \\ &= P\left(\varepsilon_t \le 0\right) = 0.5 \end{aligned}. \tag{A3}$$

Appendix B

We will analytically show that the AR-Norm model gives a larger magnitude of the error update for a higher forecast streamflow.

Firstly, we will show that the first derivate of the log-sinh transform f defined by (3) is positive and the second derivate is negative (i.e. $f'(x) > 0$ and $f''(x) < 0$) for any $b > 0$ and any x. Following some simple manipulation, we have

$$f'(x) = \frac{\cosh\left(a + bx\right)}{\sinh(a + bx)} > 0 \qquad \text{and} \qquad f''(x) = \frac{-b}{\sinh^2(a + bx)} < 0 \tag{B1}$$

Using the differentiation of inverse functions, we find the first and second derivates of the inverse transform f^{-1}

$$\left[f^{-1}\right]'(x) = \frac{1}{f'\left\{f^{-1}(x)\right\}} > 0 \quad \text{and} \quad \left[f^{-1}\right]''(x) = \frac{-f''\left\{f^{-1}(x)\right\}}{\left[f'\left\{f^{-1}(x)\right\}\right]^3} > 0, \tag{B2}$$

for any $b > 0$ and any x.

Next, we will derive the difference of magnitudes of the error update between low and high forecast streamflows. For the sake of notation simplicity, we rewrite $q = \tilde{Z}_t$ and $u = \rho\left(Z_{t-1} - \tilde{Z}_{t-1}\right)$ and assume that $u > 0$. Using Equation (4), the updated streamflow can be written as $\tilde{Q}_t^* = f^{-1}(q + u)$. The magnitude of the error update can be written as

$$\left|\tilde{Q}_t^* - \tilde{Q}_t\right| = \left|f^{-1}(q+u) - f^{-1}(q)\right| = \begin{cases} \int_0^u \left[f^{-1}\right]'(x+q)dx & \text{if} \quad u > 0 \\ \int_u^0 \left[f^{-1}\right]'(x+q)dx & \text{otherwise.} \end{cases} \tag{B3}$$

Suppose that we have two forecast streamflows $\tilde{Q}_{t,1} \le \tilde{Q}_{t,2}$ and denote the normalised forecast streamflow by $q_1 = \tilde{Z}_{t,1}$ and $q_2 = \tilde{Z}_{t,2}$ and the updated streamflow by $\tilde{Q}_{t,1}^*$ and $\tilde{Q}_{t,2}^*$. Because f is an increasing function, we have $q_1 \le q_2$. The difference in the magnitude of the error update between $\tilde{Q}_{t,1}$ and $\tilde{Q}_{t,2}$ can be derived as

$$\left|\tilde{Q}_{t,1} - \tilde{Q}_{t,1}^*\right| - \left|\tilde{Q}_{t,2} - \tilde{Q}_{t,2}^*\right| = \begin{cases} \int_0^u \left\{\left[f^{-1}\right]'(x+q_1) - \left[f^{-1}\right]'(x+q_2)\right\}dx & \text{if} \quad u > 0 \\ \int_u^0 \left\{\left[f^{-1}\right]'(x+q_1) - \left[f^{-1}\right]'(x+q_2)\right\}dx & \text{otherwise.} \end{cases} \tag{B4}$$

From (A2), we have shown that $\left[f^{-1}\right]'$ is a positive increasing function and this ensures that $\left[f^{-1}\right]'(x+q_1) - \left[f^{-1}\right]'(x+q_2) \le 0$. Finally we have

$$\left|\tilde{Q}_{t,1} - \tilde{Q}_{t,1}^*\right| \le \left|\tilde{Q}_{t,2} - \tilde{Q}_{t,2}^*\right|. \tag{B5}$$

Therefore, the error update at larger forecast streamflows is always larger than error update at lower forecast streamflows.

Acknowledgments

This work is part of the WIRADA (Water Information Research and Development Alliance) streamflow forecasting project funded under CSIRO Water for a Healthy Country Flagship. We would like to thank Durga Shrestha for valuable suggestions that led to substantial strengthening of the manuscript. We would like to thank two reviewers, Bettina Schaefli and Mark Thyer, for their careful reviews and valuable recommendations, which have improved the quality of this manuscript considerably.

Table of Tables

Table 1: Catchment characteristics.

Table 2: Comparison of the NSE calculated at (a) the receding limb and (b) the rising limb of the hydrograph for three different error models.

Table 3: Comparison of the skill scores based on CRPS and RMSEP (denoted by CRPS_SS and RMSEP_SS) for three different error models.

Table of Figures

Figure 1: Map of US (top) and Australian (bottom) catchments.

Figure 2: An example of over-correction caused by the AR-Norm model in the Mitta Mitta catchment. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating.

Figure 3: The fraction of instances where $D_t > \left|Q_{t-1} - \tilde{Q}_{t-1}\right|$ (i.e., instances where over-correction may occur in the AR-Norm model and where error updating is restricted in the RAR-Norm model) for the AR-Norm and RAR-Norm models for Australian catchments.

Figure 4: Forecast streamflows for the Orara catchment for an example 1-year period. Top panel shows streamflows forecast with AR-Norm model, bottom panel shows streamflows forecast with the RAR-Norm model. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating. Tick marks in the x-axis denote the instance of updating where $D_t > \left|Q_{t-1} - \tilde{Q}_{t-1}\right|$.

Figure 5: An example of over-correction caused by the AR-Raw model in the Mitta Mitta catchment. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating.

Figure 6: Forecast streamflows for the Abercrombie catchment for the period between 01/08/1997 and 15/09/1997. Top panel shows streamflows forecast with AR-Raw model, bottom panel shows streamflows forecast with the RAR-Norm model. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating. Gray shading denotes instances of over-correction caused by the AR-Raw model.

Figure 7: NSE of streamflows forecast with the AR-Norm, AR-Raw and RAR-Norm models (colours). Performance of the corresponding base hydrological models is shown by hatched blocks.

Figure 8: Comparison of the observed streamflows (Q_t) and forecast streamflows ($\tilde{Q}_t$), as forecast: 1) with the base hydrological model (circles), and 2) with the base hydrological model and error updating models (dots) for the Orara catchment.

Figure 9: PIT-uniform probability plots. Curves on the diagonal indicate perfectly reliable forecasts.

References

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, 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 Forecasting (SCHEF) to lead times of 9 days, J Hydrol, 10.1016/j.jhydrol.2014.08.010, 2014.

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, 10.5194/hess-17-4209-2013, 2013.

Duan, Q., Schaake, J., Andreassian, 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, 10.1016/j.jhydrol.2005.07.031, 2006.

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, 10.1016/0022-1694(94)90057-4, 1994.

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, 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, 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, 10.1002/2013WR014185, 2014.

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., Franks, S. W., and Kuczera, G.: Confronting Input Uncertainty in Environmental Modelling, in: Calibration of Watershed Models, edited by: Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N., and Turcotte, R., American Geophysical Union, Washington D.C., 49-68, 2003.

Li, M., Wang, Q. J., and Bennett, J.: Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, Water Resour Res, 49, 5913-5929, 10.1002/wrcr.20445, 2013.

Morawietz, M., Xu, C. Y., and Gottschalk, L.: Reliability of autoregressive error models as post-processors for probabilistic streamflow forecasts, Adv. Geosci., 29, 109-118, 10.5194/adgeo-29-109-2011, 2011.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J Hydrol, 10, 282-290, 10.1016/0022-1694(70)90255-6, 1970.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, J Hydrol, 279, 275-289, 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, 10.1029/2011wr010603, 2012.

Salamon, P., and Feyen, L.: Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation, Water Resour Res, 46, W12501, 10.1029/2009wr009022, 2010.

Schaefli, B., Talamba, D. B., and Musy, A.: Quantifying hydrological modeling errors through a mixture of normal distributions, J Hydrol, 332, 303-315, 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, 10.1029/2009wr008933, 2010.

Seo, D. J., Herr, H. D., and Schaake, J. C.: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction, Hydrol. Earth Syst. Sci. Discuss., 3, 1987-2035, 10.5194/hessd-3-1987-2006, 2006.

Shrestha, D. L., and Solomatine, D. P.: Data‐driven approaches for estimating uncertainty in rainfall‐runoff modelling, International Journal of River Basin Management, 6, 109-122, 10.1080/15715124.2008.9635341, 2008.

Thyer, M., Kuczera, G., and Wang, Q. J.: Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation, J Hydrol, 265, 246-257, 10.1016/S0022-1694(02)00113-0, 2002.

Vaze, J., Perraud, J. M., Teng, J., Chiew, F. H. S., Wang, B., and Yang, Z.: Catchment Water Yield Estimation Tools (CWYET), 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, Brisbane, 2011.

Wang, Q. J., and Robertson, D. E.: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, Water Resour Res, 47, W02546, 10.1029/2010WR009333, 2011.

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, 10.1029/2011WR010973, 2012.

World Meteorological Organization: Simulated real-time intercomparison of hydrological models, World Meteorological Organization, Geneva, Switzerland, 1992.

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, 10.1080/02626660209492964, 2002.

Xu, C. Y.: Statistical analysis of parameters and residuals of a conceptual water balance model - Methodology and case study, Water Resour Manag, 15, 75-92, 10.1023/A:1012559608269, 2001.

Table 1: Catchment characteristics.

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
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 2: Comparison of the NSE calculated at (a) the receding limb and (b) the rising limb of the hydrograph for three different error models.

	(a) $\tilde{Q}_t \le \tilde{Q}_{t-1}$				(b) $\tilde{Q}_t > \tilde{Q}_{t-1}$			
	Proportion of flows	AR-Norm	AR-Raw	RAR-Norm	Proportion of flows	AR-Norm	AR-Raw	RAR-Norm
Abercrombie	82%	0.11	-0.41	0.52	19%	0.58	0.66	0.65
Mitta Mitta	82%	0.95	0.91	0.95	18%	0.81	0.86	0.86
Orara	85%	0.94	0.91	0.95	15%	0.86	0.86	0.83
Tarwin	71%	0.90	0.91	0.90	29%	0.18	0.77	0.76
Amite	69%	0.76	0.82	0.84	31%	0.82	0.82	0.85
Guadalupe	83%	0.75	0.35	0.77	15%	0.24	0.55	0.45
San Marcos	82%	0.80	0.66	0.80	17%	0.63	0.64	0.64

Table 3: Comparison of the skill scores based on CRPS and RMSEP (denoted by CRPS_SS and RMSEP_SS) for three different error models.

	CRPS_SS (%)			RMSEP_SS (%)		
	AR-Norm	AR-Raw	RAR-Norm	AR-Norm	AR-Raw	RAR-Norm
Abercrombie	64.1	62.3	66.3	75.1	73.7	74.7
Mitta Mitta	80.3	79.7	80.7	84.1	83.2	84.0
Orara	74.0	75.7	75.5	81.7	80.7	81.4
Tarwin	74.9	79.3	78.8	86.1	85.1	86.1
Amite	67.5	68.3	69.5	71.0	70.9	71.2
Guadalupe	57.4	60.9	59.8	76.3	75.2	77.2
San Marcos	68.8	66.0	68.9	73.9	73.9	74.3

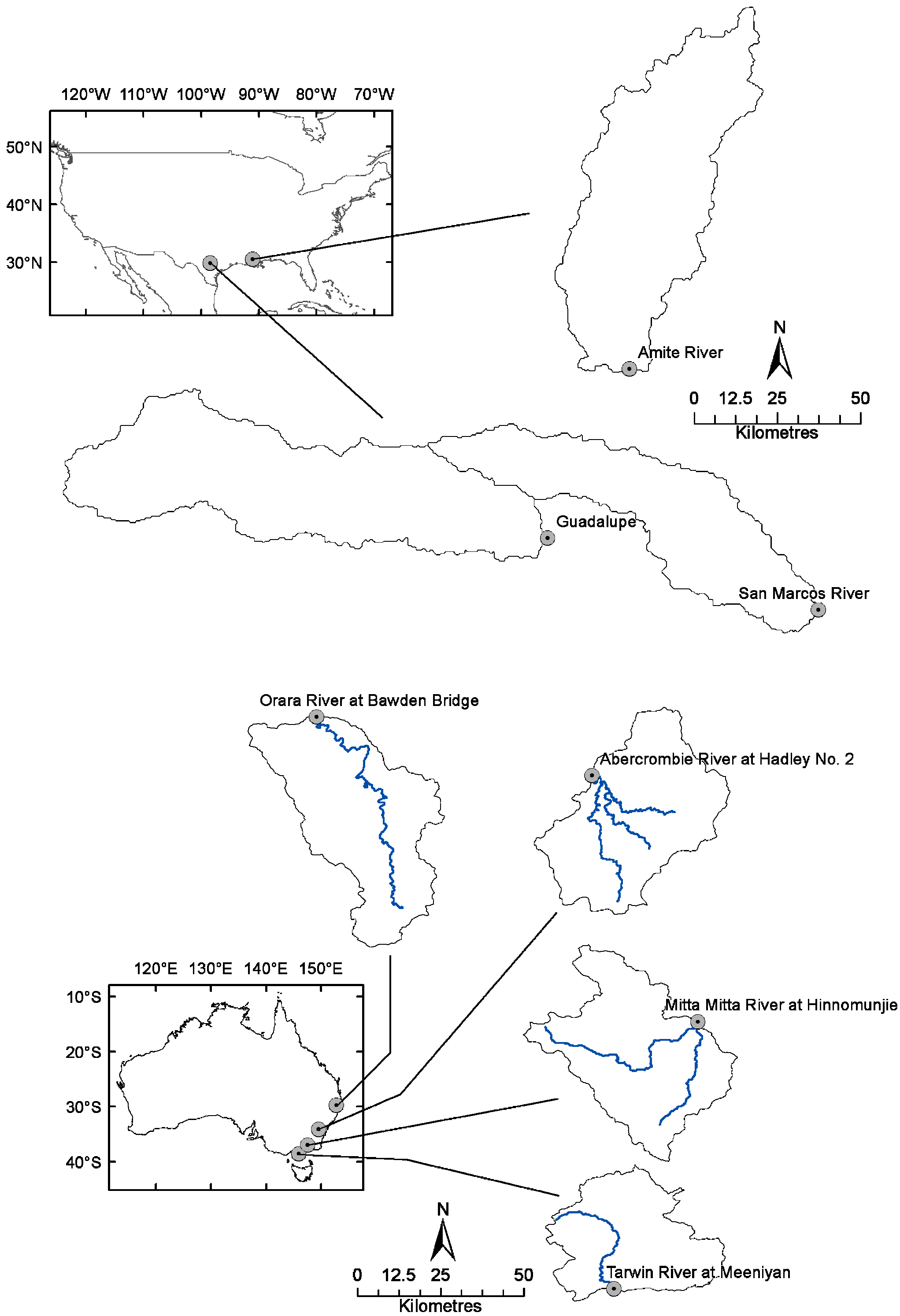

Figure 1: Map of US (top) and Australian (bottom) catchments.

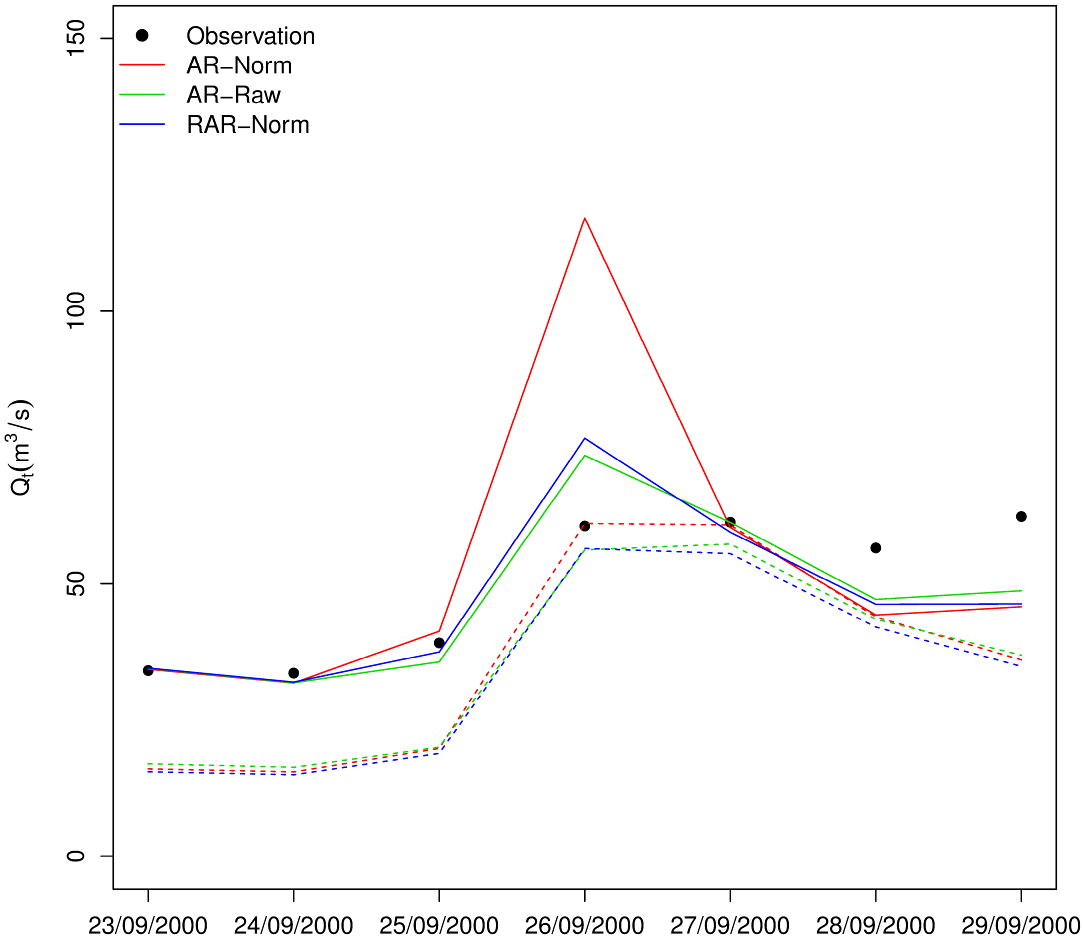

Figure 2: An example of over-correction caused by the AR-Norm model in the Mitta Mitta catchment. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating.

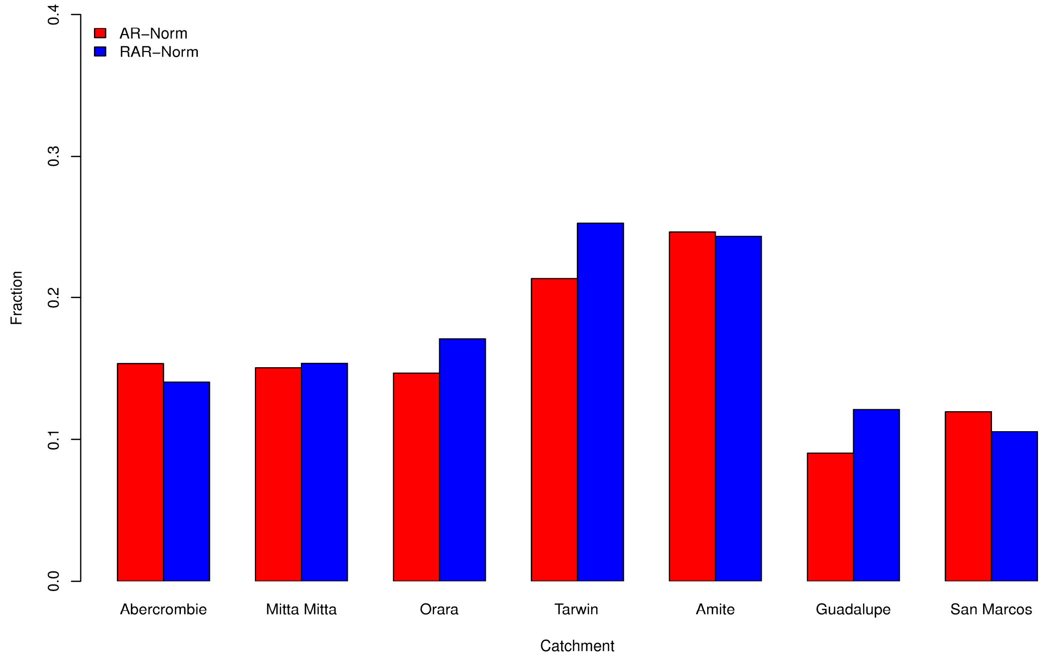

Figure 3: The fraction of instances where $D_t > \left| Q_{t-1} - \tilde{Q}_{t-1} \right|$ (i.e., instances where over-correction may occur in the AR-Norm model and where error updating is restricted in the RAR-Norm model) for the AR-Norm and RAR-Norm models for Australian catchments.

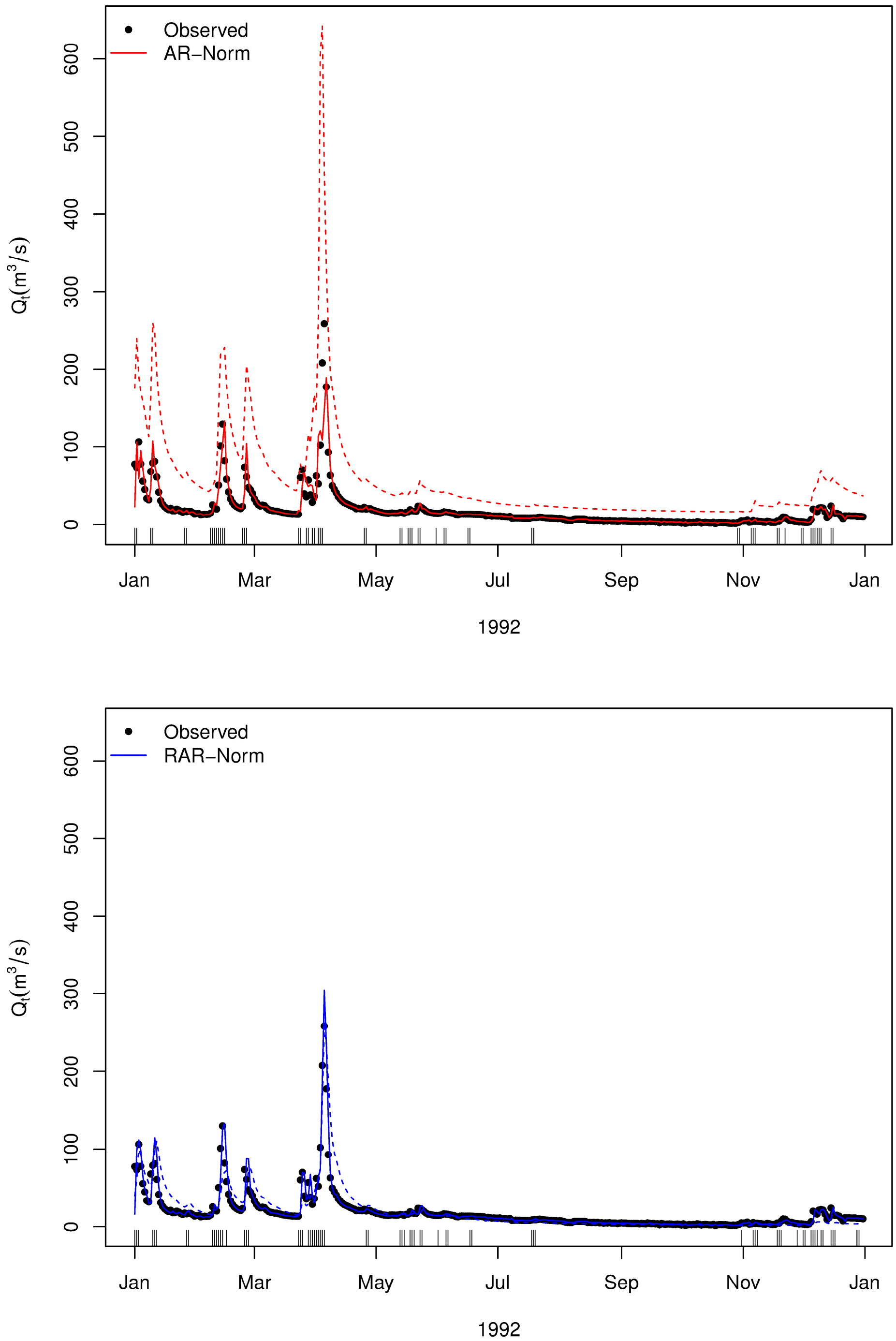

Figure 4: Forecast streamflows for the Orara catchment for an example 1-year period. Top panel shows streamflows forecast with AR-Norm model, bottom panel shows streamflows forecast with the RAR-Norm model. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating. Tick marks in the x-axis denote the instance of updating where $D_t > \left|Q_{t-1} - \tilde{Q}_{t-1}\right|$.

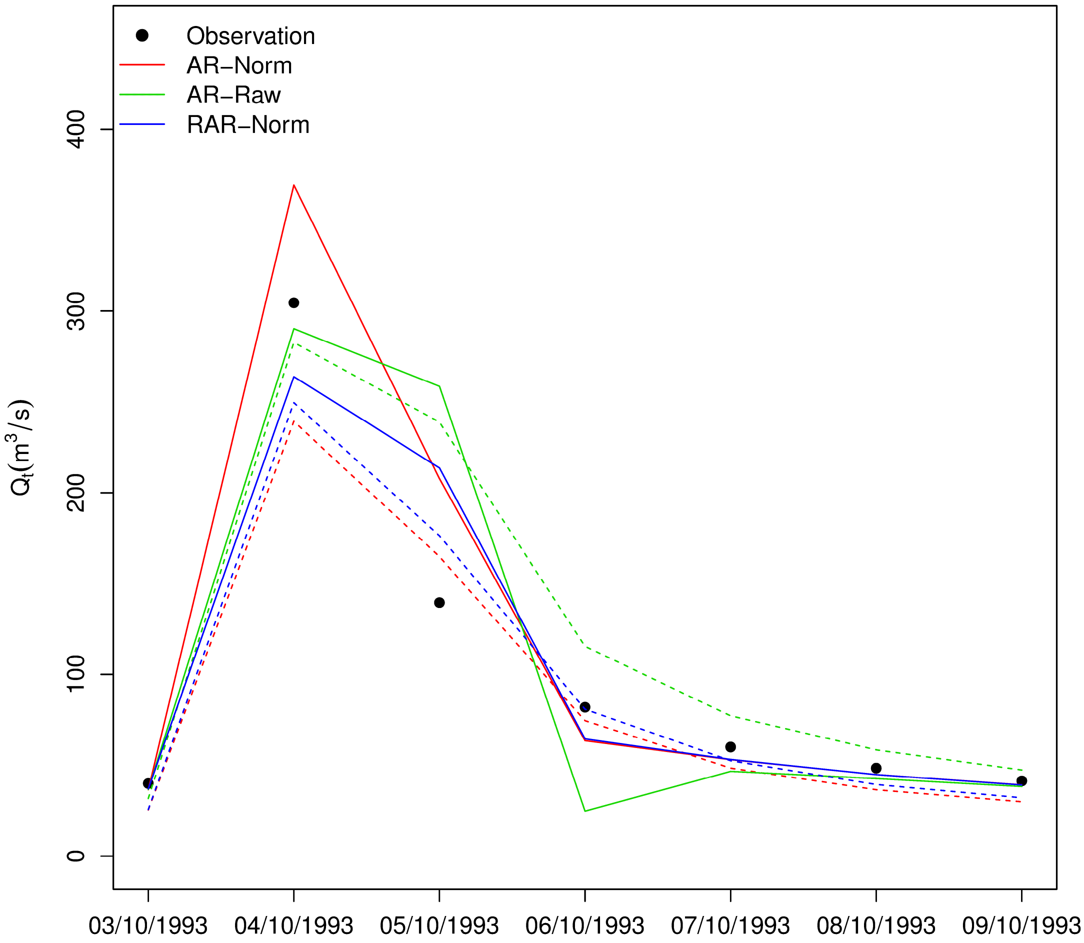

Figure 5: An example of over-correction caused by the AR-Raw model in the Mitta Mitta catchment. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating.

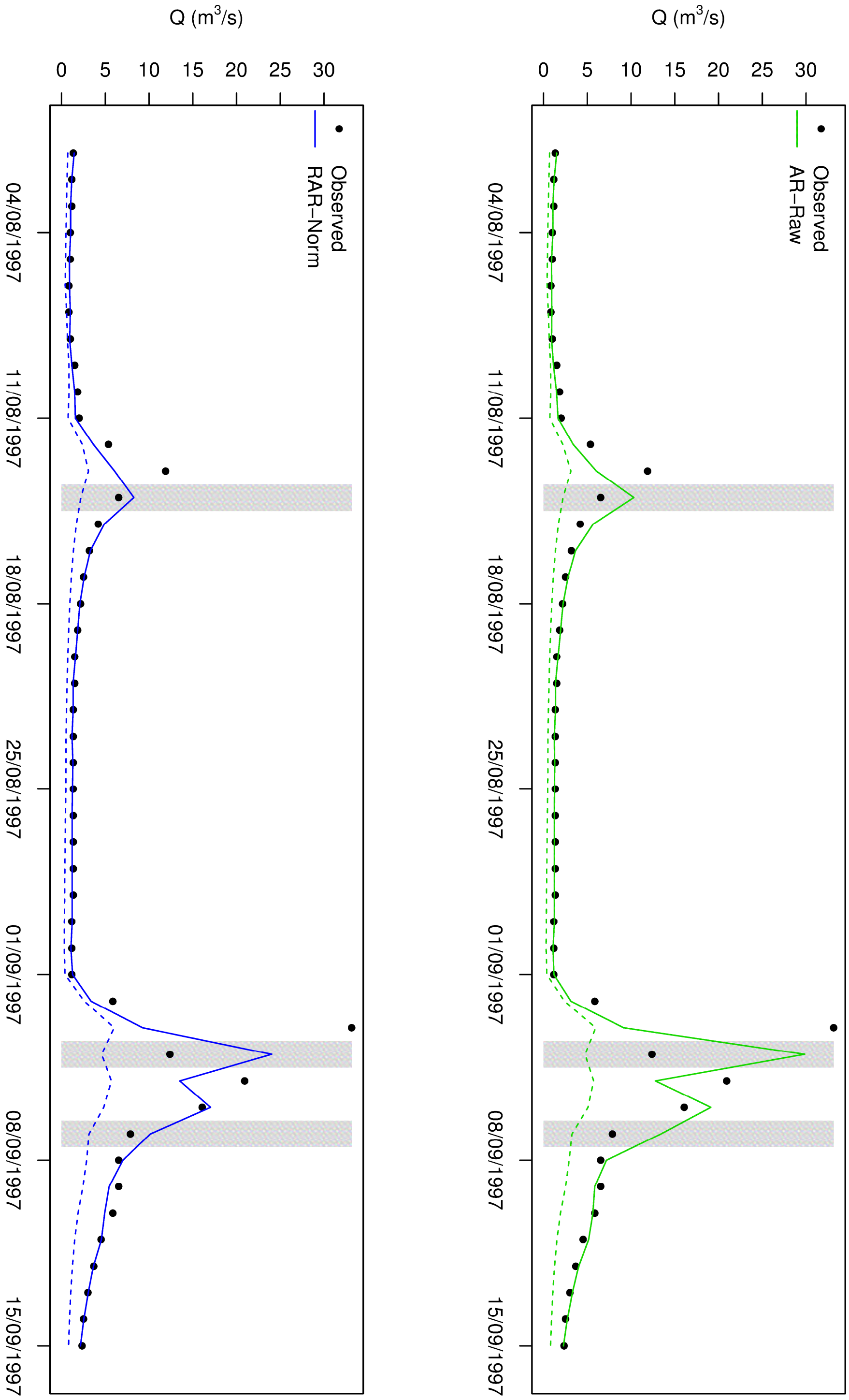

Figure 6: Forecast streamflows for the Abercrombie catchment for the period between 01/08/1997 and 15/09/1997. Top panel shows streamflows forecast with AR-Raw model, bottom panel shows streamflows forecast with the RAR-Norm model. Dashed lines: forecasts from the base hydrological model (i.e., without error updating). Solid lines: forecasts with error updating. Gray shading denotes instances of over-correction caused by the AR-Raw model.

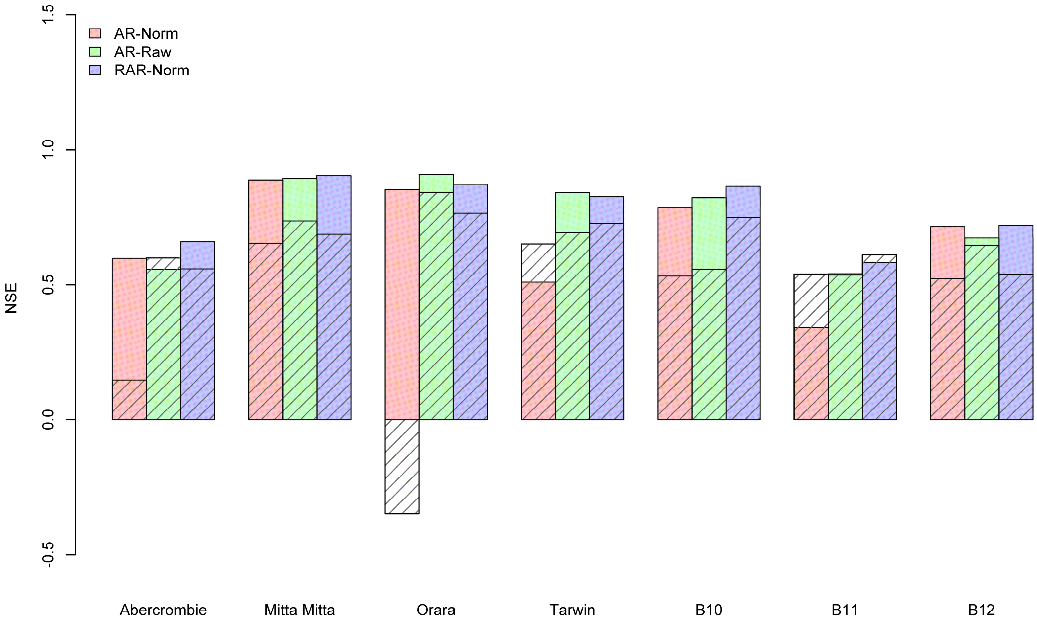

Figure 7: NSE of streamflows forecast with the AR-Norm, AR-Raw and RAR-Norm models (colours). Performance of the corresponding base hydrological models is shown by hatched blocks.

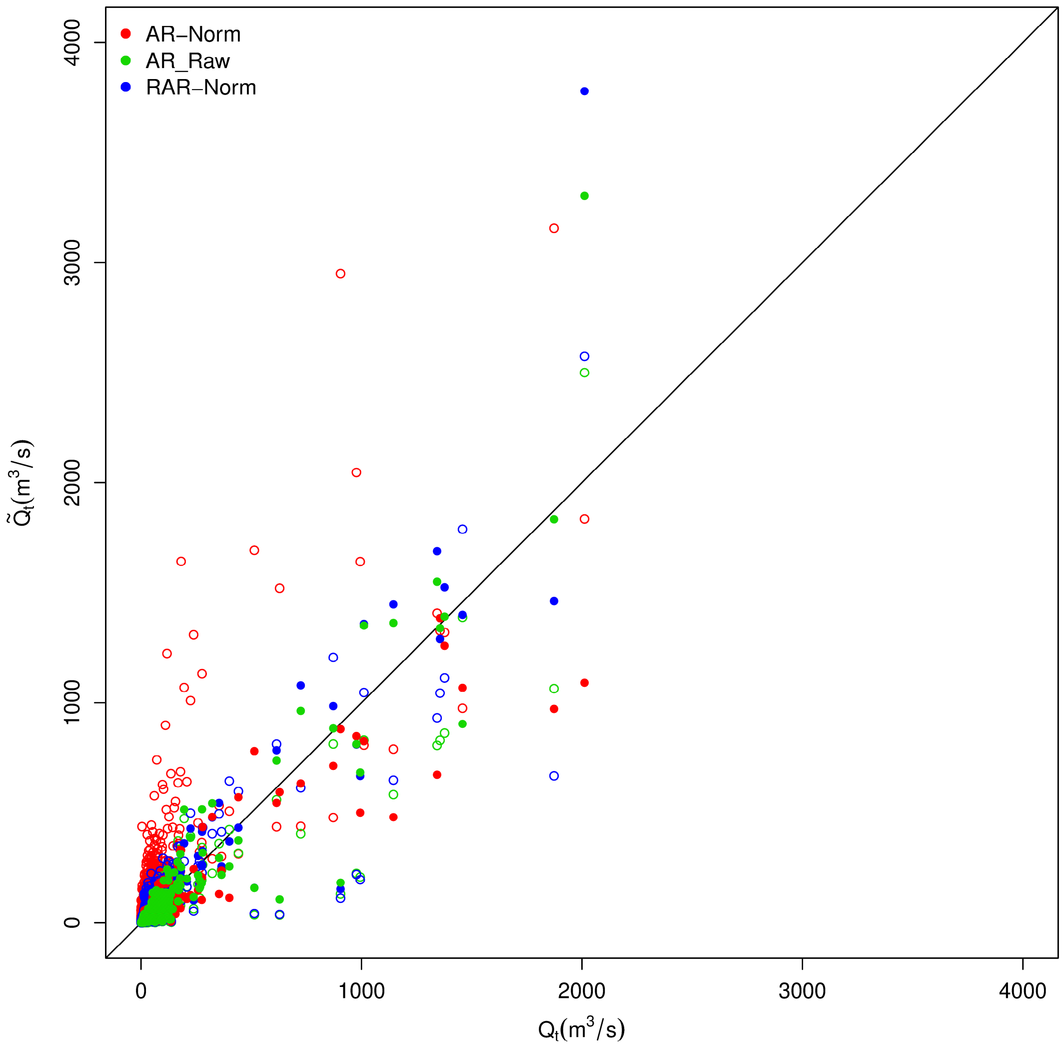

Figure 8: Comparison of the observed streamflows (Q_t) and forecast streamflows ($\tilde{Q}_t$), as forecast: 1) with the base hydrological model (circles), and 2) with the base hydrological model and error updating models (dots) for the Orara catchment.

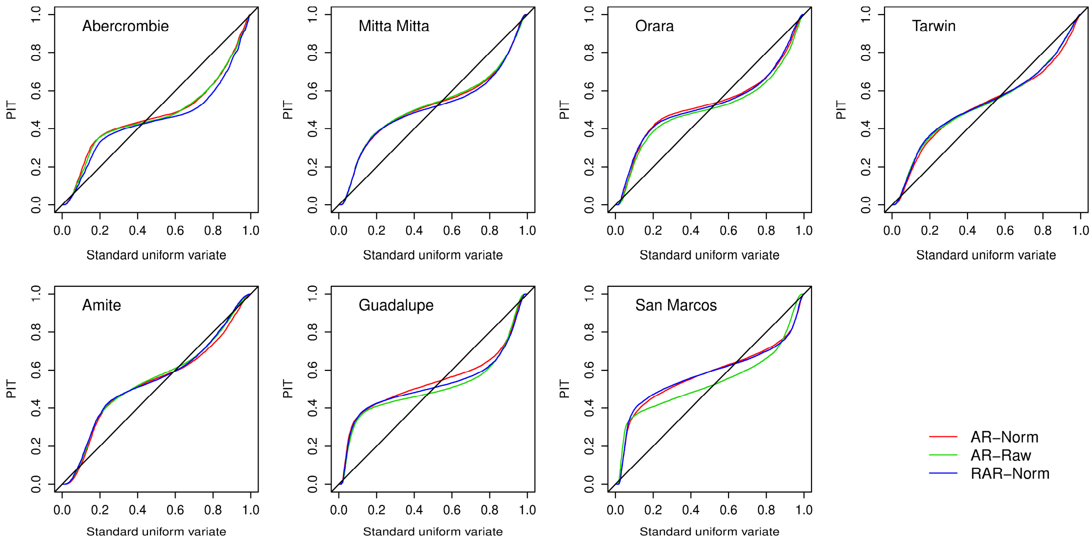

Figure 9: PIT-uniform probability plots. Curves on the diagonal indicate perfectly reliable forecasts.