Jason Hunter, Mark Thyer, David McInerney, Dmitri Kavetski

High-quality probabilistic predictions for existing hydrological models with common objective functions

Proceedings of the Hydrology and Water Resources Symposium (HWRS 2021), 2021, pp.626-639

ISBN number for HWRS 2021 is 978-1-925627-53-4

High-quality probabilistic predictions for existing hydrological models with common objective functions

Jason Hunter, Mark Thyer, David McInerney, Dmitri Kavetski

School of Civil, Environmental and Mining Engineering, University of Adelaide

Email: jason.hunter@adelaide.edu.au

ABSTRACT

Probabilistic predictions describe the uncertainty in modelled streamflow, which is a critical input for many environmental modelling applications. A residual error model typically produces the probabilistic predictions in tandem with a hydrological model that predicts the deterministic streamflow. However, many objective functions that are commonly used to calibrate the parameters of the hydrological model make (implicit) assumptions about the errors that do not match the properties (e.g. of heteroscedasticity and skewness) of those errors. The consequence of these assumptions is often low-quality probabilistic predictions of errors, which reduces the practical utility of probabilistic modelling. Our study has two aims:

Firstly, to evaluate the impact of objective function inconsistency on the quality of probabilistic predictions;

Secondly, to demonstrate how a simple enhancement to a residual error model can rectify the issues identified with inconsistent objective functions in Aim 1, and thereby improve probabilistic predictions in a wide range of scenarios.

Our findings show that the enhanced error model enables high-quality probabilistic predictions to be obtained for a range of catchments and objective functions, without requiring any changes to the hydrological modelling or calibration process. This advance has practical benefits that are aimed at increasing the uptake of probabilistic predictions in real-world applications, in that the methods are applicable to existing hydrological models that are already calibrated, simple to implement, easy to use and fast. Finally, these methods are available as an open-source R-shiny application and an R-package function.

1 INTRODUCTION

Daily streamflow predictions produced by hydrological models are widely used in practical environmental and water resource applications, but contain inherent uncertainties as a consequence of simplifying natural systems (Beven 1989; Oreskes et al. 1994). Increasingly, hydrologists are expected to report the uncertainty in the streamflow predictions along with the predictions themselves (Welsh et al. 2013; Ball et al. 2016), which can pose a difficulty in that a majority of existing hydrological models provide only determinstic 'point' predictions of streamflow and do not describe the predictive uncertainty (see Moore 1985; Perrin et al. 2003; Brunner & Simmons 2012 for examples of well-known hydrological models that do not directly consider uncertainty). Probabilistic predictions from residual error models can describe predictive uncertainty, e.g. by providing more information on the likely range of streamflow, and can thereby avoid the false sense of security associated with point streamflow predictions (Biondi et al. 2010; Farmer & Vogel 2016).

However, there are several current barriers to the routine inclusion of probabilistic predictions in practical streamflow modelling, such as a perceived complexity in their generation and difficulties in their interpretation (Pappenberger & Beven 2006). One such barrier investigated in this study is the impact of the objective function used to calibrate the hydrological model on the quality of the probabilistic predictions. Objective functions for calibrating deterministic predictions are typically chosen for reasons other than facilitating high-quality probabilistic predictions, and many common objective functions leave the model residuals poorly-described.

A common approach for probabilistic uncertainty quantification is the residual error model approach, where a probabilistic error model (the residual error model) is added to the deterministic predictions of a hydrological model to represent the combined contribution of multiple sources of errors (Sorooshian et al. 1983; Schoups & Vrugt 2010; Evin et al. 2014). Practical implementations of the residual error model approach typically follow the 'post-processor' strategy, where the hydrological model parameters are estimated first using an objective function, followed by a separate estimation of residual error model parameters (Engeland et al. 2010; Evin et al. 2014; Li et al. 2016; McInerney et al. 2018). This post-processor approach is particularly attractive for practical applications because, at least in principle, it enables probabilistic predictions to be generated using hydrological models calibrated with user-specified objective functions. The post-processor approach is interesting because the quality of probabilistic predictions using this methodology depends on both the residual error model *and* the objective function as two separate processes. The objective function produces residuals with specific structures and patterns (i.e. makes 'assumptions' about the residuals) that are separately interpreted by the residual error model.

Substantial previous research has identified robust residual error models that produce high-quality probabilistic predictions (Kuczera 1983b; Smith & Marshall 2010; Wang et al. 2012; Del Giudice et al. 2013; Cheng et al. 2014; McInerney et al. 2017). These robust residual error models provide a realistic description of the statistical properties of the errors in the deterministic predictions. Common statistical properties of residual errors include (but are not limited to) their mean, heteroscedasticity (i.e. larger errors in larger flows), asymmetry (skewness) and temporal persistence (i.e. multiple consecutive errors with the same sign and similar magnitude) (Sorooshian & Dracup 1980; Bates & Campbell 2001; Evin et al. 2013; Smith et al. 2015; Sun et al. 2017). Residual error models found to perform well at the daily scale include those that transform streamflow using the Box-Cox transformation (McInerney et al. 2017).

This study uses the term 'objective function inconsistency' specifically to describe the modelling scenario in which the assumptions about the mean, heteroscedasticity and skewness made by the objective function are different from the assumptions about the mean, heteroscedasticity and skewness made by the residual error model. A typical example of 'objective function inconsistency' relevant to our study and broader practice arises if the hydrological model parameters are calibrated using an objective function without streamflow transformations, but predictive uncertainty is estimated using a residual error model that applies a Box-Cox transformation (which assumes that the errors are heteroscedastic and skewed).

These considerations lead to the first question, '*Do differences/inconsistencies between the assumptions of the objective function and the assumptions of the residual error model impact on the quality of the probabilistic predictions*?' This question is by no means trivial, but to our knowledge there is no study that has examined this issue. However, studies have evaluated the performance of inconsistent objective functions without comparison (Evin et al. 2014; Li et al. 2016), or have examined the performance of post-processed approaches with purely *consistent* scenarios (McInerney et al. 2018).

Previous studies provide insights into modifications to the residual error model that can assist with predictive quality in scenarios of objective function inconsistency. For example, the Error Reduction and Representation In Stages (ERRIS) approach developed by Li et al. (2016) included a flow-dependent bias correction to the residual error model, as part of its four-stage model for day-ahead forecasts. In another study, Jiang et al. (2019) explored time-varying representations of the mean, variance and distributional form of residual errors. These two studies demonstrate that modifications to the mean parameter of the residual error model warrant investigation, which leads to a new question: '*How robust are modifications or enhancements to the residual error model across a wide range of catchments and a wide range of hydrological model objective functions?'*

This study has the following aims:

1. Evaluate the impact of objective function inconsistency on the quality of probabilistic predictions;
2. Demonstrate how a simple enhancement to a residual error model can rectify the issues identified with inconsistent objective functions in Aim 1, and thereby improve probabilistic predictions in a wide range of scenarios.

A broader objective of this work is to facilitate the uptake of probabilistic predictions by researchers and practitioners in hydrology and water resources. Hence, there is an emphasis on simple and practical modelling approaches that can be incorporated with relatively minor effort into existing and future applications.

The paper is organised as follows. Section 2 outlines the theory of probabilistic models used in this work. Section 3 describes the case study data and methods, including the selection of catchments, hydrological model, objective functions, performance metrics and residual error diagnostics. Section 4 reports the case study results. Section 5 discusses the findings and provides recommendations for current applications and future research, and Section 6 summarises the key conclusions of the study.

2 THEORY

The models and methods used in this work are based on those developed in McInerney et al. (2018), wherein the residual error model is calibrated in a different process from the hydrological model. The hydrological model produces a single time-series of predictions $q_t^{\boldsymbol{\theta}_h}$ where $\boldsymbol{\theta}_h$ represents the parameters used to calibrate the hydrological model and q_t is the streamflow predictions at time t. The residual error model produces multiple probabilistic predictions Q_t that represent multiple possible scenarios, which are formulated by adding the hydrological model predictions $q_t^{\boldsymbol{\theta}_h}$ to modelled residual errors η_t.

$$z\left(Q_t;\boldsymbol{\theta}_z\right) = z\left(q_t^{\boldsymbol{\theta}_h};\boldsymbol{\theta}_z\right) + \eta_t \tag{1}$$

$z\left(Q_t;\boldsymbol{\theta}_z\right)$ and $z\left(q_t^{\boldsymbol{\theta}_h};\boldsymbol{\theta}_z\right)$ represent the residual model predictions Q_t and the hydrological model predictions $q_t^{\boldsymbol{\theta}_h}$ (respectively) that are transformed. Data transformations are commonly applied to streamflow prior to calibration: this helps remove heteroscedasticity and skewness from the predictions. In this study we use the Box-Cox data transformation (Box & Cox 1964)

$$z\left(q;\boldsymbol{\theta}_z\right) = \begin{cases} \dfrac{\left(q+A\right)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log\left(q+A\right) & \text{if } \lambda = 0 \end{cases} \tag{2}$$

with parameters $\boldsymbol{\theta}_z = \left\{\lambda, A\right\}$, where λ is a power parameter A is a shift (or offset) parameter, q is streamflow predictions and log is the natural logarithm function. The transformation parameters $\boldsymbol{\theta}_z$ in the residual error model are fixed prior to calibration at $\lambda = 0.2$ and $A = 0$, in accordance with recommendations made in earlier works (McInerney et al. 2017; Jiang et al. 2019). The residual errors η_t are modelled with a first-order autoregressive (AR1) model, which uses parameters $\mathbf{\theta}_\eta = \{\mu, \phi, \sigma\}$. The mean parameter μ_t may be variable in time.

$$\eta_t = \mu_t + \phi(\eta_{t-1} - \mu_{t-1}) + N(0, \sigma^2) \tag{3}$$

The residual error model in McInerney et al. (2018) assumes a zero-mean parameter, or $\mu = 0$. Here we introduce an enhanced residual error model which uses a linear-mean parameter instead, that is conditioned on the simulated streamflow. We therefore compare the performance of both residual error models: a reference model with a zero-mean and an enhanced model with a linear mean.

Reference residual error model

The mean is assumed to be zero: that is; the uncertainty about the simulated streamflow predictions (in transformed space) is Gaussian and not offset.

$$\mu_t = 0 \tag{4}$$

Where μ_t is the mean parameter at each timestep. This is a common assumption made in residual error modelling (e.g. Kuczera 1983a; Evin et al. 2014; McInerney et al. 2017; Sun et al. 2017; McInerney et al. 2018).

Enhanced residual error model

The mean is assumed to follow a linear function that is based on the transformed streamflow predictions from the hydrologic model.

$$\mu_t = \alpha + \beta \cdot z\left(q_t^{\mathbf{\theta}_h}; \mathbf{\theta}_z\right) \tag{5}$$

Where α represents the intercept, and β represents the slope. This linear mean parameter is not especially common, although it has been explored elsewhere as part of more complex methods in forecasting studies (Li et al. 2016; Jiang et al. 2019). For complete details of the models and methods used, including generation of the probabilistic predictions and both the enhanced and reference residual error models, the reader is referred to Hunter et al. (2021).

3 CASE STUDY MATERIAL AND METHODS

Catchments and observed data

The case study considers 54 Australian catchments with a (relatively) diverse range of hydroclimatology (Figure 1). Daily rainfall, potential evapotranspiration (PET) and streamflow time series are obtained from the Australian Bureau of Meteorology Hydrologic Reference Stations (HRS) dataset (http://www.bom.gov.au/water/hrs). For each catchment, 10 years of continuous observed data is selected. Ephemeral catchments are excluded from this study because they require more complex residual error modelling methods (Smith et al. 2010; McInerney et al. 2019; Wang et al. 2020). The selected time periods vary between catchments, with the earliest starting in 1970 and the latest ending in 2012.

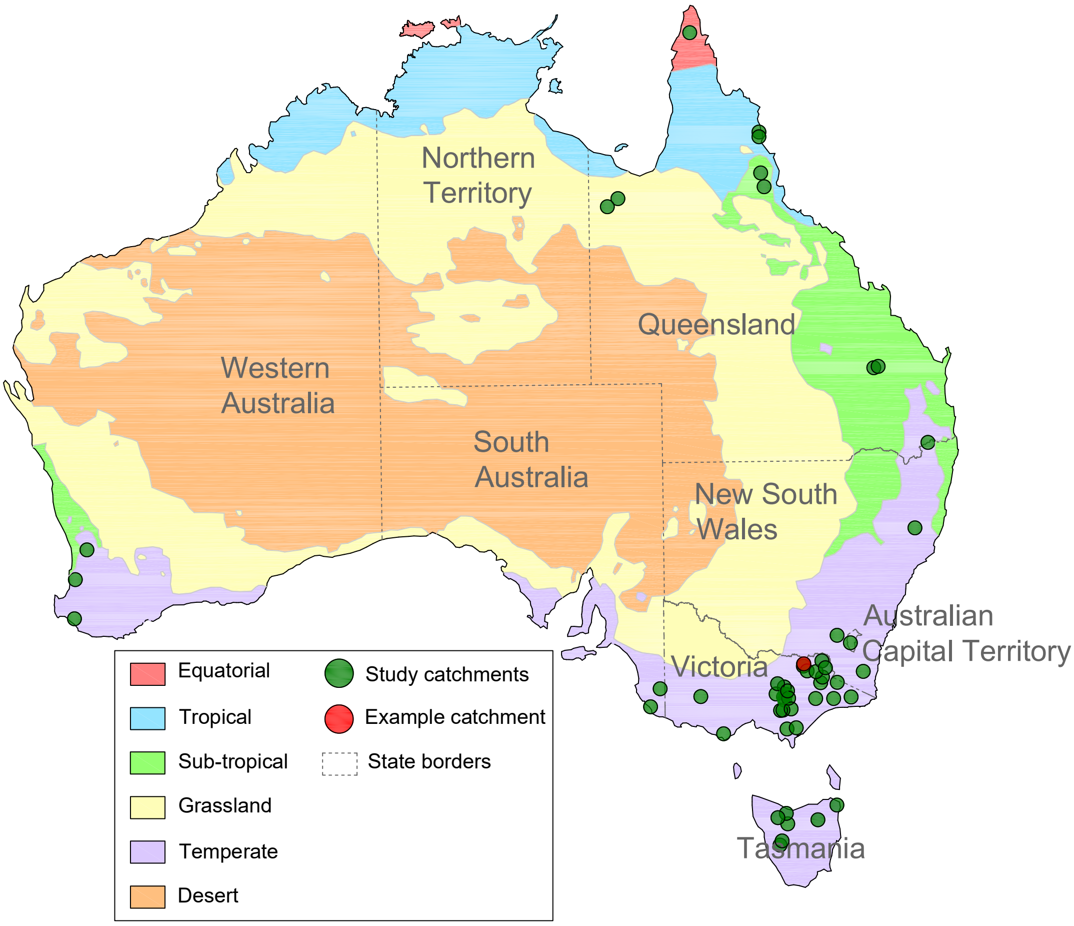

Figure 1: Locations of the 54 case study catchments in Australia. The catchment represented in the diagnostics of Aim 1 is circled in red. The archetypal Köppen classifications (Stern et al. 2000) are shown.

Hydrological model

The hydrological model used in this study is GR4J, a daily conceptual rainfall-runoff model with two storages and four parameters (Perrin et al. 2003). GR4J has been tested under a range of hydroclimatic conditions and is used in the streamflow forecasting services of France and Australia (Oudin et al. 2006; Ficchì et al. 2016; Woldemeskel et al. 2018; Lerat et al. 2020). The four parameters of GR4J are estimated by optimising the objective functions listed in Section 0 below, with the algorithm in the *airGR* R-package (Coron et al. 2017; Lerat et al. 2020).

Objective functions

Nine hydrological objective functions are considered and are listed in Table 1. They are all (except for the Pearson's correlation coefficient, which is included for demonstration purposes only) commonly-used in hydrological practice. When referring to these objective functions (or the model that includes them), we generally refer to the abbreviated term in the second column of Table 1. Where the same objective function structure is used (e.g. NSE structure), the objective functions differ in the parameters of the transformation in Eq. (2) that is applied to the streamflow data prior to calibration of the hydrological model. This difference in transformation parameters ensures that the calibrated hydrological parameters are distinct, and therefore these objective functions are functionally unique.

Table 1: List of objective functions used to calibrate the hydrological parameters $\boldsymbol{\theta}_h$ in this study.

Objective function structure	Abbreviation	Box-Cox λ
Nash-Sutcliffe Efficiency (Nash & Sutcliffe 1970)	NSE	1
	NSE-BC02	0.2
	NSE-BC05	0.5
	NSE-Log	0
Kling-Gupta Efficiency (Gupta et al. 2009)	KGE	1
	KGE-BC02	0.2
	KGE-BC05	0.5
Nash-Sutcliffe Efficiency and Bias (Vaze et al. 2010)	NSE-BIAS	1
Pearson Correlation Coefficient	R2	1

Note that the transformation parameter λ applied to the calibration data of the hydrological objective function in Table 1 can be different from the transformation parameter used for the residual error model, which is always $\lambda = 0.2$.

Terminology

We define a 'scheme' as a 'combination of a hydrological model and a residual error model' (Hunter et al. 2021). The term 'model' refers to either the 'hydrologic model', or 'residual error model' (both explained in more detail in Section 0).

The schemes used in this paper can be described as either *benchmark*, *baseline* or *enhanced*, where:

- *Benchmark schemes* make assumptions about the residuals that are consistent between the objective function used to calibrate the hydrological model parameters $\boldsymbol{\theta}_h$ and the residual error model. Because all of our residual error models use a Box-Cox power parameter value of $\lambda = 0.2$, the model that uses a NSE-BC02 objective function and a zero-mean assumption (reference residual error model) is our benchmark scheme.
- *Baseline schemes* can make inconsistent assumptions between the objective function and the residual error model, but still use the reference residual error model. Therefore, all schemes that use the reference residual error model are considered baseline schemes, except for the scheme that uses the NSE-BC02 objective function (which is specifically a benchmark scheme).
- *Enhanced schemes* make inconsistent assumptions between the objective function and residual error model, but are distinct from the baseline schemes in that enhanced schemes use the enhanced residual error model. All schemes that use the enhanced residual error model are enhanced schemes.

Performance evaluation and metrics

The quality of the probabilistic predictions are evaluated using three common goodness-of-fit metrics (e.g. see Renard et al. 2010; McInerney et al. 2017; Hunter et al. 2021 for details). For all metrics, a lower value indicates better performance.

- *Reliability* indicates the likelihood that the observed data might have been drawn from the probabilistic predictions, and is based on a quantification of the predictive quantile-quantile (PQQ) plot.
- *Precision* refers to the spread/width of uncertainty within the predictive distribution.
- *Volumetric bias* is a measure of how well the predictions represent the long-term water balance.

4 RESULTS

Impact of inconsistency on predictive quality

Figure 2 provides the outputs from four diagnostic tools on predictions from one example catchment (gauge 402204 in Yackandandah Creek, VIC).

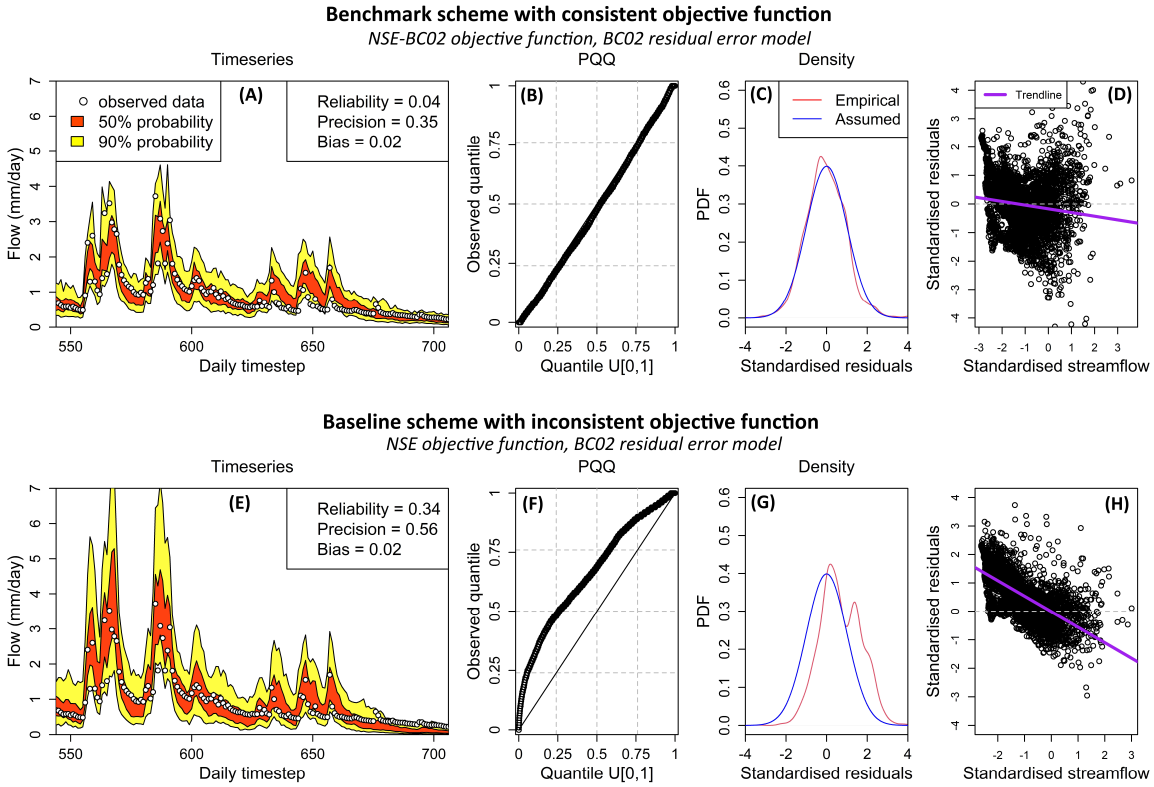

Figure 2: Diagnostics that evaluate the predictive quality and behavior of two schemes: The benchmark scheme with consistent assumptions between the (NSE-BC02) objective functions and the (BC02) residual error model (top), and a baseline scheme with inconsistent assumptions between the (NSE) objective function and the (BC02) residual error model.

Panels (A) and (E) show part of the predictive time series, of the benchmark scheme (top) and the baseline scheme (bottom) for catchment 402204. The predictive bounds of the baseline time series (Panel E) are unneccessarily large: bounds that are much larger than the observed data risks over-estimating the model uncertainty. This behaviour translates into higher values of the reliability and precision in Panel (E) as opposed to Panel (A): recalling that lower values of both metrics indicate better predictions.

Panels (B) and (F) are the predictive quantile-quantile (PQQ) plots for both benchmark (top) and baseline (bottom) schemes. A reliable scheme will have the observed data on the uniform 1:1 line (i.e. like Panel B). Panel (F) indicates a consistant underestimation of observed data, in that the 1:1 line falls below the observed data for almost all data points.

Panels (C) and (G) represent the probability density function of the standardised residuals. The residual error model assumes a Guassian distribution of the residuals (from transformed streamflow quantities), so ideally the standardised residuals (empirical distribution) will fit to the assumed Guassian distribution. This is largely the case for Panel (C), the benchmark scheme with consistent assumptions between the hydrologic model and the residual error model, but is less so for Panel (G) where the assumptions are inconsistent.

Finally, Panels (D) and (H) plot the standardised residuals (from transformed streamflow quantities) against the standardised (transformed) streamflow predictions. Recalling that both benchmark and baseline scheme assume a residual mean of zero, we can surmise that an ideal trendline would lie horizontal at zero. It is horizontal for neither panel (i.e. there is some systematic linear behaviour in the residuals), but this linear trend is far more pronounced in the inconsistent scheme, in Panel (H) than in Panel (D).

Analysis of these diagnostics therefore indicate that:

- The predictive performance of the baseline schemes are worse than the predictive performance of the benchmark scheme, because of the inconsistency in assumptions.

- There is a systematic issue with the predictive performance of the baseline schemes (e.g. Panels (F) and (H)).
- There is a linear trend in the residuals, that becomes more pronounced in the inconsistent scheme (Panel (H)).

Enhanced scheme demonstration

This section compares the predictive performance of the baseline, benchmark and enhanced schemes, thus addressing Study Aim 2.

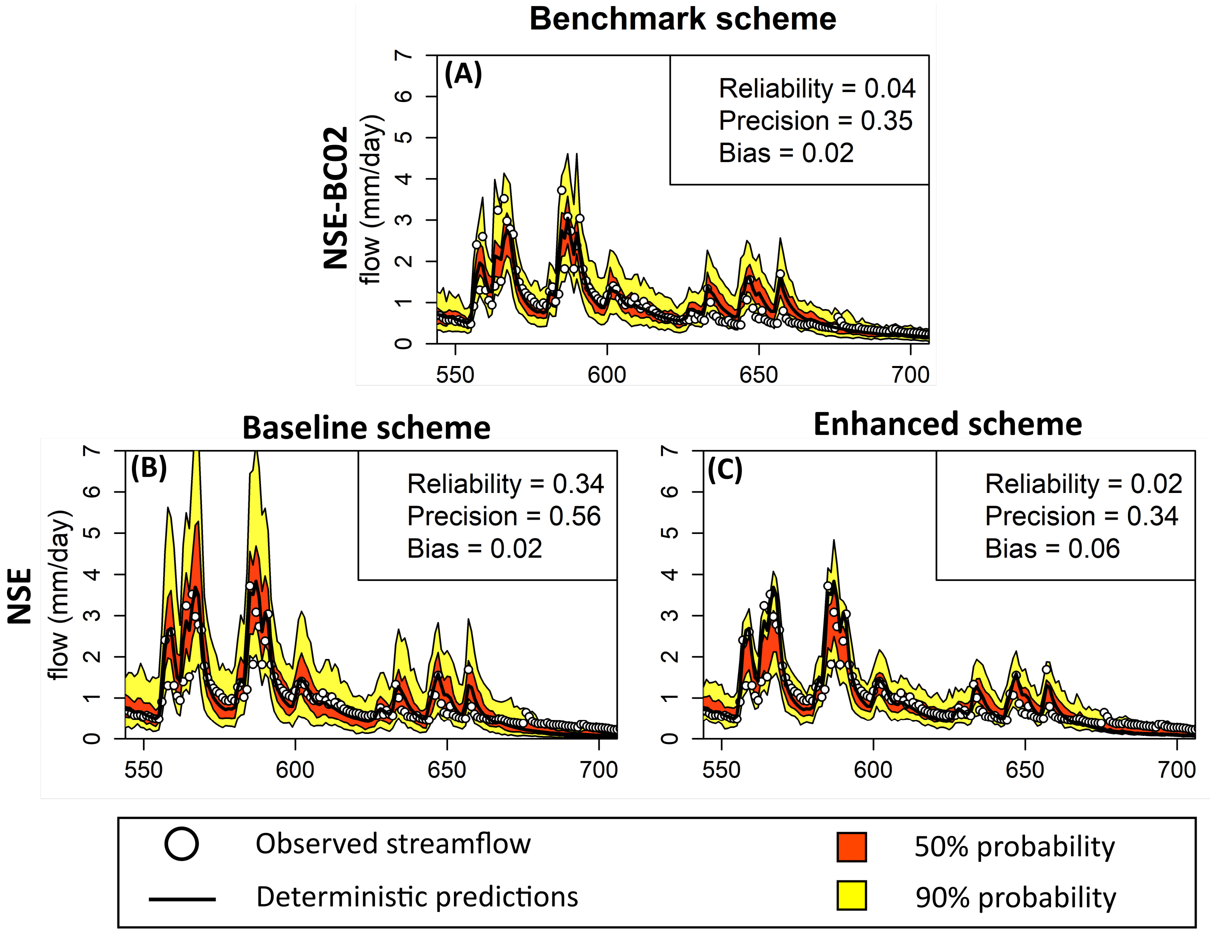

Figure 3: Comparative timeseries between (A) the benchmark scheme, (B) a baseline scheme with the NSE objective function, and (C) an enhanced scheme with the NSE objective function. Example catchment is gauge 402204 at Yackandandah Creek in VIC.

Panels (A) and (B) of Figure 3 match Panels (A) and (E) of Figure 2; and exhibit the same trend of the inconsistent baseline scheme over-estimating the observed data and thereby providing predictions that are less reliable and less precise than the benchmark scheme in Panel (A). We stress that the difference between Panels (A) and (B) are solely the objective function: Panel (A) uses a NSE-BC02 objective function with assumptions that are consistent with the BC02 residual error model, while Panel (B) uses an NSE objective function with assumptions that are different from the assumptions of the BC02 residual error model.

Panel (C) uses the same objective function as Panel (B), but is able to resolve the issues of precision and reliability Panel (B) because it uses a linear mean parameter rather than a zero-mean parameter. The reliability, precision and bias of Panel (C) is similar to the same metrics of Panel (A), the benchmark scheme – despite Panel (C) still having inconsistent assumptions between the objective function and the residual error model. Visually, the predictions in Panel (C) are closer to the observed streamflow predictions and do not overestimate the model uncertainty as much as the baseline scheme does.

Metrics for all catchments

Reliability, precision and bias are calculated for all 54 case-study catchments and objective functions, and plotted in boxplots.

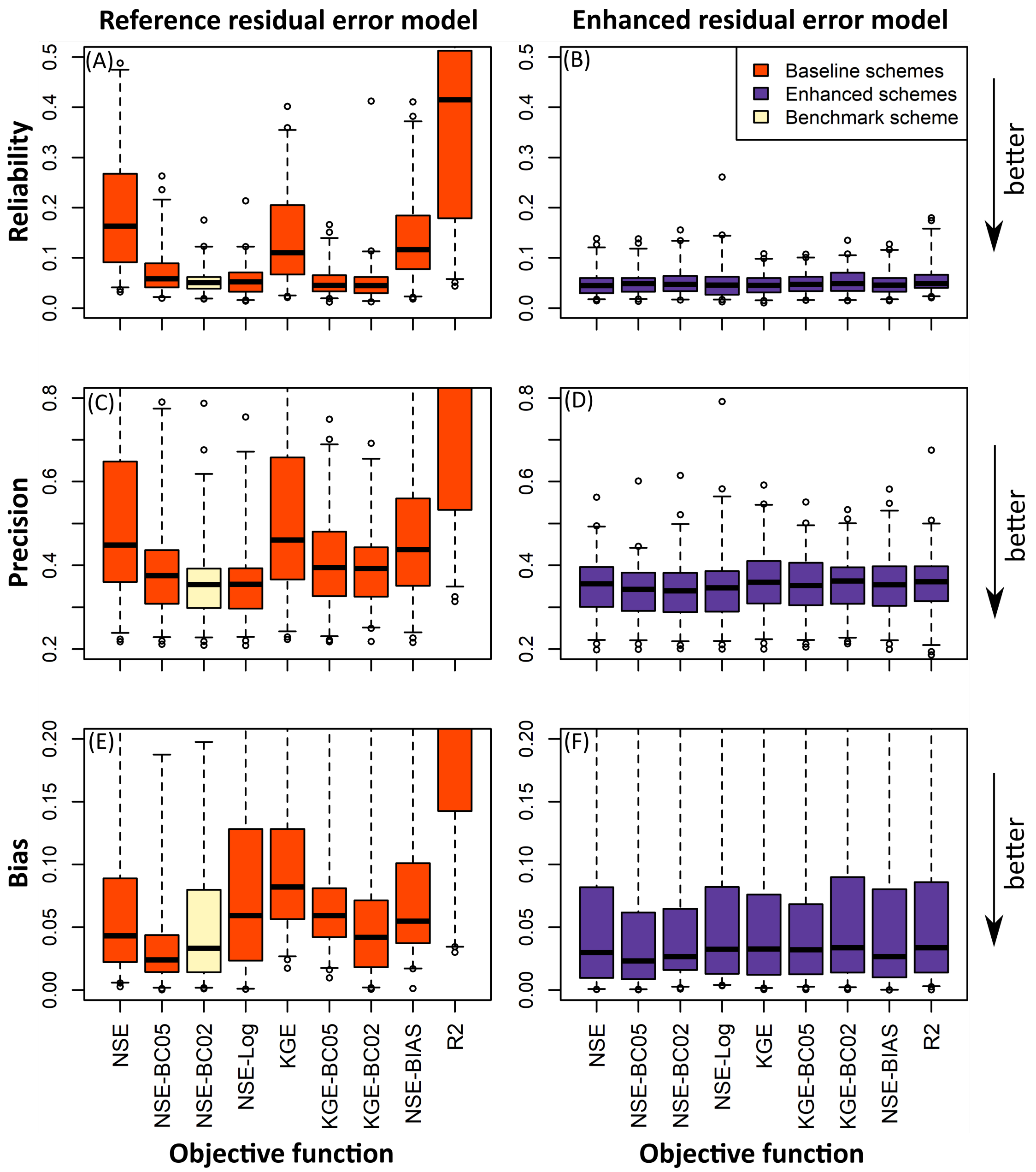

Figure 4: Metrics for 54 case-study catchments for the benchmark and baseline schemes (Panels A, C, E) and the enhanced schemes (Panels B, D, F).

Reliability

Figure 4(A) and (B) compares the reliability of 54 catchments and 9 objective functions. The benchmark scheme is one of four schemes with the best reliability, indicated by a low median reliability metric of 0.05. The baseline NSE-Log, KGE-BC02 and KGE-BC05 schemes achieve a similar reliability, whereas the baseline NSE, KGE, NSE-BIAS and R2 schemes are far less reliable (i.e. they have higher values of reliability).

In contrast to the worst-performing baseline schemes (i.e. R2, NSE-BIAS, KGE, NSE), the enhanced schemes show a large improvement in reliability. The median reliabilities for the enhanced schemes are similar to the median reliability of the benchmark scheme and range from 0.04 to 0.05 across all objective functions and case study catchments. The enhanced scheme outputs seem independent of the objective function: that is, the reliability is relatively uniform across all enhanced schemes.

Precision

Figure 4(C) and (D) compares the precision for the same 54 study catchments and 9 objective functions. The benchmark scheme is one of two schemes with the best precision, with a median precision metric value of 0.35. The precision values for the remaining baseline schemes vary widely from 0.35 (NSE-Log) to 0.93 (R2). The enhanced schemes demonstrate less variability and are approximately as precise as the benchmark scheme.

Volumetric Bias

Figure 4(E) and (F) reports the volumetric bias. The benchmark scheme is, once again, one of the top performers, with a median bias metric of 0.03. The baseline schemes perform generally worse, with median bias metrics that range from 0.04 (NSE and KGE-BC02 schemes) to 0.08 (KGE scheme), and up to 0.28 for the R2 schemes. The enhanced schemes achieve large improvements in the bias metric (relative to the baseline schemes with the same objective function), and perform similar to the benchmark scheme.

Summary of case study results

These results from all the case-study catchments reinforce the findings from the example catchment diagnostics. That is;

- The baseline schemes, with inconsistent objective functions and a zero-mean assumption, generally shows a degredation of predictive performance from the benchmark scheme with a consistent objective function.
- This issue in performance is resolved when the enhanced residual error model, with a linear-mean assumption, is used – despite there still being inconsistent assumptions between the objective function and the residual error model.

5 DISCUSSION

Most of the baseline schemes exhibit poor performance relative to the benchmark scheme. The only difference between baseline and benchmark schemes is the objective function: Therefore, the deterioration in quality must be due to this influence. The residual diagnostics indicate that the change in streamflow transformation between the objective function and the residual error model, which is done to treat residual heteroscedasticity and skewness, introduces a linear structure into the model residuals (Figure 2H). Empirically, this behaviour is systematic across all catchments and objective functions.

A valuable finding is that the enhanced schemes produce results that are equitable with the consistent benchmark scheme, and also with each other regardless of objective function. Each objective function produces residuals with different structures and assumptions: This finding indicates that the linear mean is sufficiently flexible to describe sets of residuals that are structurally diverse.

Practical benefits

The study findings indicate that an enhanced residual error model can be used with a wide range of objective functions to provide a high quality probabilistic predictions. This achievement offers three major practical benefits:

Benefit 1: Method can be applied to existing hydrological models that are already calibrated

This saves time and effort for practioners, because there is no need to recalibrate the hydrological model to facilitate high-quality probabilistic predictions. Changing or re-calibrating the objective function to one that is consistent in assumptions with a residual error model is often impractically time-consuming, especially with more complex distributed models that are slower to calibrate. More generally, practitioners may favour certain objective functions, or even develop their own, based on previous experience with their specific hydrological models, catchments and/or operational objectives, while researchers may be interested in specific objective functions as part of their study scope.

Replacing the residual error model with one that makes assumptions that are the same as the hydrological objective function (i.e. use a scheme similar to the benchmark scheme that we have used in this study) might appear to be a reasonable option, but doing this is often ill-advised, for a few reasons:

- The statistical assumptions underlying an objective function can be opaque and/or unknown, especially for composite objective functions that combine several other objective functions, e.g. NSE-BIAS.
- Creating and testing customised residual error models for such objective functions can be challenging and risks producing low-quality probabilistic predictions.
- Many residual error models that have consistent assumptions with common objective functions are known to produce poor-quality probabilistic predictions. For example, the standard least-squares residual error model is consistent with the NSE objective function with no data transformation, but cannot manage data heteroscedasticity or skewness so produces poor-quality predictions (Sorooshian & Dracup 1980; Kuczera 1983a; Schoups & Vrugt 2010; Cheng et al. 2014; McInerney et al. 2017).

Benefit 2: Approach is simple to implement, easy to use and fast

The method proposed in this work are considered *simple* to implement because it only uses two inputs: the time series of observed streamflow, and the time series of the hydrological model predictions. This takes advantage of the post-processor strategy for residual error modelling, where the paramters for the residual error model are calibrated in an entirely separate stage from the parameters of the hydrologic model. The post-processor approach is generally more flexible than traditional joint approaches where the parameters of both the hydrological model and residual error model are estimated simultaneously. Post-processor strategies also enable a wider range of parameter estimation methods for residual error modelling: the Method-of-Moments is used here, but Maximum Likelihood and full Bayesian methods can also be used (Evin et al. 2014; Li et al. 2016; Jiang et al. 2019).

The calibration of the residual error model uses the Method-of-Moments for parameter estimation, which is both fast and easy to use while still providing parameter estimates and predictive quality similar to more complex statistical inference methods (McInerney et al. 2018). The Method-of-Moments does not require a likelihood function or optimisation, and therefore requires less specialised statistical knowledge and fewer computational tools.

Benefit 3: Algorithms are available in R-shiny Web app and R Functions

As a demonstration of the simpliticy of use, the methods presented in this paper are available on github as open-source software (https://github.com/Jasenter/Probabilistic_App), where they can be interfaced as either an R-function or an R-shiny application (both of these interfaces are provided in the github repository).

Future work

The following research questions are noted for further investigation:

- *Generality of findings across broader applications*. The conclusions of this study are derived empirically from the catchments, hydrological model and objective functions tested: this research could benefit from experimentation with a more diverse set of catchments (including those affected by ephemerality or snowmelt), and more hydrological models or objective functions.
- *Further development of residual error models and/or objective functions.* The applicability of the flow-dependent mean enhancement to more complex residual error models and/or objective functions, such as those that account for seasonality, zero flows or extreme flows (Wang & Robertson 2011; Li et al. 2013; Liu et al. 2020), warrants investigation.

6 CONCLUSIONS

High-quality probabilistic predictions of streamflow are useful in practice and in research. Our work has demonstrated that there is an issue in the predictive quality of probabilistic predictions when the assumptions of the objective function used to calibrate the hydrological model are inconsistent with the assumptions of the residual error model. There is limited scope to adjust the residual error model, because it must be able to account for residual heteroscedasticity and skewness and not every residual error model can perform this task equally. Our work addresses a research gap related to combining residual error models with inconsistent objective functions in a post-processor context, where the two processes are separated and therefore can hold inconsistent assumptions about the residuals. We show that there is an issue with predictive quality when the assumptions are inconsistent, and that this issue can be resolved with an enhanced residual model that replaces the traditional zero-mean parameter with a linear mean. We demonstrate the efficacy of our enhanced residual error model on 54 catchments and 9 objective functions, and provide complete details of the model construction in Hunter et al. (2021).

Acknowledgements

This research is supported by an Australian Government Research Training Program (RTP) Scholarship. Catchment data was obtained from the Hydrologic Reference Stations database provided by the Australian Bureau of Meteorology (http://www.bom.gov.au/water/hrs). Simulations were performed on the Phoenix cluster hosted by The University of Adelaide. This work was conducted on the traditional lands of the Kaurna people, and we acknowledge their custodianship and pay our respects to their elders past, present and future. We also acknowledge the traditional custodians of the catchments and rivers that are used in this study.

7 REFERENCES

Ball, J, Babister, M, Nathan, R, Weeks, W, Weinmann, E, Retallick, M & Testoni, I 2016, Australian Rainfall and Runoff: A guide to flood estimation, *Commonwealth of Australia (Geoscience Australia)*.

Bates, BC & Campbell, EP 2001, A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, vol. 37, no. 4, pp. 937-947, DOI 10.1029/2000WR900363.

Beven, K 1989, Changing ideas in hydrology — The case of physically-based models, *Journal of Hydrology*, vol. 105, no. 1, pp. 157-172, DOI 10.1016/0022-1694(89)90101-7.

Biondi, D, Versace, P & Sirangelo, B 2010, Uncertainty assessment through a precipitation dependent hydrologic uncertainty processor: An application to a small catchment in southern Italy, *Journal of Hydrology*, vol. 386, no. 1, pp. 38-54, DOI 10.1016/j.jhydrol.2010.03.004.

Box, GE & Cox, DR 1964, An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211-252, DOI 10.1111/j.2517-6161.1964.tb00553.x.

Brunner, P & Simmons, CT 2012, HydroGeoSphere: A Fully Integrated, Physically Based Hydrological Model, *Ground Water*, vol. 50, no. 2, pp. 170-176, DOI 10.1111/j.1745-6584.2011.00882.x.

Cheng, Q-B, Chen, X, Xu, C-Y, Reinhardt-Imjela, C & Schulte, A 2014, Improvement and comparison of likelihood functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme, *Journal of Hydrology*, vol. 519, Part B, pp. 2202-2214, DOI 10.1016/j.jhydrol.2014.10.008.

Coron, L, Thirel, G, Delaigue, O, Perrin, C & Andréassian, V 2017, The suite of lumped GR hydrological models in an R package, *Environmental Modelling & Software*, vol. 94, pp. 166-171, DOI 10.1016/j.envsoft.2017.05.002.

Del Giudice, D, Honti, M, Scheidegger, A, Albert, C, Reichert, P & Rieckermann, J 2013, Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrol. Earth Syst. Sci.*, vol. 17, no. 10, pp. 4209-4225, DOI 10.5194/hess-17-4209-2013.

Engeland, K, Renard, B, Steinsland, I & Kolberg, S 2010, Evaluation of statistical models for forecast errors from the HBV model, *Journal of Hydrology*, vol. 384, no. 1-2, pp. 142-155, DOI 10.1016/j.jhydrol.2010.01.018.

Evin, G, Kavetski, D, Thyer, M & Kuczera, G 2013, Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, *Water Resources Research*, vol. 49, no. 7, pp. 4518-4524, DOI 10.1002/wrcr.20284.

Evin, G, Thyer, M, Kavetski, D, McInerney, D & Kuczera, G 2014, Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resources Research*, vol. 50, no. 3, pp. 2350-2375, DOI 10.1002/2013WR014185.

Farmer, WH & Vogel, RM 2016, On the deterministic and stochastic use of hydrologic models, *Water Resources Research*, vol. 52, no. 7, pp. 5619-5633, DOI 10.1002/2016WR019129.

Ficchì, A, Perrin, C & Andréassian, V 2016, Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, *Journal of Hydrology (Amsterdam)*, vol. 538, pp. 454-470, DOI 10.1016/j.jhydrol.2016.04.016.

Gupta, H, Kling, H, Yilmaz, KK & Martinez, GF 2009, Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, vol. 377, no. 1–2, pp. 80-91, DOI 10.1016/j.jhydrol.2009.08.003.

Hunter, J, Thyer, M, McInerney, D & Kavetski, D 2021, Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions, *Journal of Hydrology*, p. 126578, DOI 10.1016/j.jhydrol.2021.126578.

Jiang, X, Gupta, HV, Liang, Z & Li, B 2019, Towards Improved Probabilistic Predictions for Flood Forecasts Generated using Deterministic Models, *Water Resources Research*, vol. 55, no. 11, pp. 9519-9543, DOI 10.1029/2019WR025477.

Kuczera, G 1983a, Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, *Water Resources Research*, vol. 19, no. 5, pp. 1151-1162, DOI 10.1029/WR019i005p01151.

Kuczera, G 1983b, Improved parameter inference in catchment models: 2. Combining different kinds of hydrologic data and testing their compatibility, *Water Resources Research*, vol. 19, no. 5, pp. 1163-1172, DOI 10.1029/WR019i005p01163.

Lerat, J, Thyer, M, McInerney, D, Kavetski, D, Woldemeskel, F, Pickett-Heaps, C, Shin, D & Feikema, P 2020, A robust approach for calibrating a daily rainfall-runoff model to monthly streamflow data, *Journal of Hydrology*, vol. 591, p. 125129, DOI 10.1016/j.jhydrol.2020.125129.

Li, M, Wang, Q & Bennett, J 2013, Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, *Water Resources Research*, vol. 49, no. 9, pp. 5913-5929, DOI 10.1002/wrcr.20445.

Li, M, Wang, Q, Bennett, J & Robertson, D 2016, Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting, *Hydrol. Earth Syst. Sci.*, vol. 20, no. 9, pp. 3561-3579, DOI 10.5194/hess-20-3561-2016.

Liu, L, Wang, QJ & Xu, Y-P 2020, Temporally varied error modelling for improving simulations and quantifying uncertainty, *Journal of Hydrology*, vol. 586, p. 124914, DOI 10.1016/j.jhydrol.2020.124914.

McInerney, D, Kavetski, D, Thyer, M, Lerat, J & Kuczera, G 2019, Benefits of explicit treatment of zero flows in probabilistic hydrological modelling of ephemeral catchments, *Water Resources Research*, vol. 55, no. 12, pp. 11035-11060, DOI 10.1029/2018WR024148.

McInerney, D, Thyer, M, Kavetski, D, Bennett, B, Lerat, J, Gibbs, M & Kuczera, G 2018, A simplified approach to produce probabilistic hydrological model predictions, *Environmental Modelling & Software*, vol. 109, pp. 306-314, DOI 10.1016/j.envsoft.2018.07.001.

McInerney, D, Thyer, M, Kavetski, D, Lerat, J & Kuczera, G 2017, Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, *Water Resources Research*, vol. 53, no. 3, pp. 2199-2239, DOI 10.1002/2016WR019168.

Moore, RJ 1985, The probability-distributed principle and runoff production at point and basin scales, *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, vol. 30, no. 2, pp. 273-297, DOI 10.1080/02626668509490989.

Nash, JE & Sutcliffe, JV 1970, River flow forecasting through conceptual models: 1. A discussion of principles., *Journal of Hydrology*, vol. 10, pp. 257-274, DOI 10.1016/0022-1694(70)90255-6.

Oreskes, N, Shrader-Frechette, K & Belitz, K 1994, Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, *Science*, vol. 263, no. 5147, pp. 641-646.

Oudin, L, Andreassian, V, Mathevet, T, Perrin, C & Michel, C 2006, Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, *Water Resources Research*, vol. 42, no. 7, p. W07410, DOI 10.1029/2005WR004636.

Pappenberger, F & Beven, KJ 2006, Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resources Research*, vol. 42, no. 5, p. W05302, DOI 10.1029/2005WR004820.

Perrin, C, Michel, C & Andreassian, V 2003, Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, vol. 279, no. 1-4, pp. 275-289, DOI 10.1016/S0022-1694(03)00225-7.

Renard, B, Kavetski, D, Thyer, M, Kuczera, G & Franks, SW 2010, Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resources Research*, vol. 46, no. 5, p. W05521, DOI 10.1029/2009WR008328
Schoups, G & Vrugt, JA 2010, A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors, *Water Resources Research*, vol. 46, no. 10, p. W10531, DOI 10.1029/2009WR008933.
Smith, T & Marshall, L 2010, Exploring uncertainty and model predictive performance concepts via a modular snowmelt-runoff modeling framework, *Environmental Modelling and Software*, vol. 25, no. 6, pp. 691-701, DOI 10.1016/j.envsoft.2009.11.010.
Smith, T, Marshall, L & Sharma, A 2015, Modeling residual hydrologic errors with Bayesian inference, *Journal of Hydrology*, vol. 528, pp. 29-37, DOI 10.1016/j.jhydrol.2015.05.051.
Smith, T, Sharma, A, Marshall, L, Mehrotra, R & Sisson, S 2010, Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resources Research*, vol. 46, no. 12, p. W12551, DOI 10.1029/2010WR009514.
Sorooshian, S & Dracup, JA 1980, Stochastic parameter estimation procedures for hydrological rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, vol. 16, no. 2, pp. 430-442, DOI 10.1029/WR016i002p00430.
Sorooshian, S, Gupta, H & Fulton, J 1983, Evaluation of Maximum Likelihood Parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility, *Water Resources Research*, vol. 19, no. 1, pp. 251-259, DOI 10.1029/WR019i001p00251.
Stern, H, De Hoedt, G & Ernst, J 2000, Objective classification of Australian climates, *Australian Meteorological Magazine*, vol. 49, no. 2, pp. 87-96.
Sun, R, Yuan, H & Liu, X 2017, Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty estimation, *Journal of Hydrology*, vol. 554, pp. 680-692, DOI 10.1016/j.jhydrol.2017.09.041.
Vaze, J, Post, DA, Chiew, FHS, Perraud, JM, Viney, NR & Teng, J 2010, Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, *Journal of Hydrology*, vol. 394, no. 3–4, pp. 447-457, DOI 10.1016/j.jhydrol.2010.09.018.
Wang, QJ, Bennett, JC, Robertson, DE & Li, M 2020, A Data Censoring Approach for Predictive Error Modeling of Flow in Ephemeral Rivers, *Water Resources Research*, vol. 56, no. 1, p. e2019WR026128, DOI 10.1029/2019WR026128.
Wang, QJ & Robertson, DE 2011, Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences, *Water Resources Research*, vol. 47, no. 2, p. W02546, DOI 10.1029/2010WR009333.
Wang, QJ, Shrestha, DL, Robertson, DE & Pokhrel, P 2012, A log-sinh transformation for data normalization and variance stabilization, *Water Resources Research*, vol. 48, no. 5, p. W05514, DOI 10.1029/2011WR010973.
Welsh, WD, Vaze, J, Dutta, D, Rassam, D, Rahman, JM, Jolly, ID, Wallbrink, P, Podger, GM, Bethune, M, Hardy, MJ, Teng, J & Lerat, J 2013, An integrated modelling framework for regulated river systems, *Environmental Modelling & Software*, vol. 39, pp. 81-102, DOI 10.1016/j.envsoft.2012.02.022.
Woldemeskel, F, McInerney, D, Lerat, J, Thyer, M, Kavetski, D, Shin, D, Tuteja, N & Kuczera, G 2018, Evaluating residual error approaches for post-processing monthly and seasonal streamflow forecasts, *Hydrol. Earth Syst. Sci. Discuss.*, vol. 22, no. 12, pp. 1-40, DOI 10.5194/hess-2018-214.