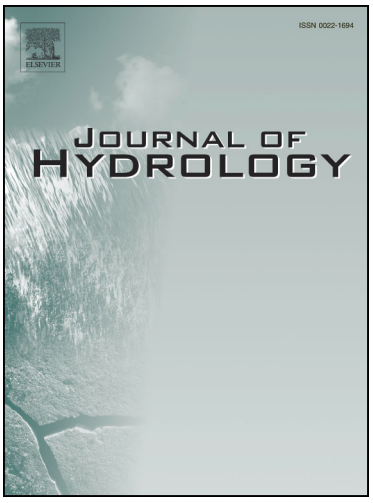
Journal Pre-proofs



Research papers

A multi-model evaluation of probabilistic streamflow predictions via residual error modelling

Jonathan Romero-Cuellar, Rezgar Arabzadeh, James R. Craig, Bryan A. Tolson, Juliane Mai

PII: S0022-1694(24)00547-X
DOI: https://doi.org/10.1016/j.jhydrol.2024.131152
Reference: HYDROL 131152

To appear in: *Journal of Hydrology*

Received Date: 13 July 2023
Revised Date: 8 December 2023
Accepted Date: 27 December 2023

Please cite this article as: Romero-Cuellar, J., Arabzadeh, R., Craig, J.R., Tolson, B.A., Mai, J., A multi-model evaluation of probabilistic streamflow predictions via residual error modelling, *Journal of Hydrology* (2024), doi: https://doi.org/10.1016/j.jhydrol.2024.131152

A Multi-model Evaluation of Probabilistic Streamflow Predictions via Residual Error Modelling

Jonathan Romero-Cuellar[1], Rezgar Arabzadeh[1], James R. Craig[1], Bryan A. Tolson[1], and Juliane Mai[2,3]

[1] Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada.

[2] Computational Hydrosystems, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Saxony, Germany.

[3] Center for Scalable Data Analytics and Artificial Intelligence - ScaDS.AI, Leipzig, Saxony, Germany.

Corresponding author: Jonathan Romero-Cuellar (jromeroc@uwaterloo.ca)

- **Keywords:** Uncertainty analysis; Probabilistic prediction; Residual error; Streamflow; Hydrological modeling; Postprocessing method

Abstract

Probabilistic streamflow predictions are valuable tools for predictive uncertainty estimation, hydrologic risk management, and support for decision-making in water resources. Usually, predictive uncertainty quantification is developed and assessed using only a single hydrological model, making it difficult to generalize to other model configurations. To tackle this issue, we assess changes in the model performance ranking of diverse streamflow models by applying a residual error model post-processing approach to multiple basins and multiple models. This assessment employed 141 basins from the Great Lakes watershed covering the USA and Canada, and 13 diverse streamflow models, which are evaluated using deterministic and probabilistic performance metrics. As the first study to implement probabilistic methods to diverse streamflow models applied to a multitude of basins, the analysis here examines the dependence of probabilistic streamflow estimation quality on model quality. Our findings show that streamflow model choice influences the robustness of probabilistic predictions. It was found that moving from deterministic to probabilistic predictions using a post-processing approach does not change the streamflow model performance ranking for the best and worst deterministic models, but models of intermediate rank in deterministic evaluation do not have consistent ranking when evaluated in probabilistic mode. Post-processing residual errors of long short-term memory (LSTM) network models are consistently the best-performing model in terms of deterministic and probabilistic metrics. This study highlights the significance of combining deterministic streamflow model predictions with residual error models for improving the quality and increasing the value of hydrological predictions, quantifying uncertainty, and facilitating decision-making in operational water management. It also clarifies the degree to which probabilistic predictions depend upon good model performance and can compensate for poor model performance.

1 Introduction

Although deterministic model predictions offer information to support hydrologic planning, design, and decision-making in environmental and water resource applications, they do not provide uncertainty estimates needed for hydrologic risk management. Deterministic models omit the model uncertainty associated with simulated responses. Throughout this paper, the term "deterministic" model prediction will refer to models that provide only one value of streamflow at any given time while probabilistic refer to models that provide estimates of streamflow in the form of a probability density. These model predictions suffer from many sources of uncertainty, including input data uncertainty, parameter uncertainty, and model structural uncertainty (Moges et al., 2021; Gupta & Govindaraju, 2023). Uncertainty quantification has become an integral part of risk assessment, and hydrological models have been used to provide information for policymakers who are increasingly demanding that model results include estimates of uncertainty (Bastola et al., 2012; Dias et al., 2020). Unfortunately, many of the most widely used models in hydrology (e.g., SWAT, GR4J, HBV, or VIC) are deterministic and provide only a single trace of streamflow output (point predictions). In contrast, probabilistic predictions provide the uncertainty estimation associated with a dynamic basin system by generating an ensemble comprised of multiple sets of streamflow predictions (Farmer & Vogel, 2016). Given the benefits of uncertainty characterization, it is important to convert deterministic predictions to probabilistic ones that can support hydrological risk management (e.g., Vogel, 2017; Reggiani et al., 2022; Shabestanipour et al., 2023).

Deterministic predictions from a given hydrological model can be converted to probabilistic predictions using a post-processing strategy to add a residual error model (REM) to the deterministic predictions (e.g., Evin et al., 2014). In general, the post-processing strategy is implemented in two steps. First, hydrological model parameters are estimated using an objective function and calibration algorithm. Second, a REM form is selected, and the parameters of REM are inferred separately. REMs are statistical models to represent the relationship between model outputs (e.g., streamflow) and observations (Ye et al., 2014). Because they characterize the expected difference between model output and observation, REMs include the merged impact of several sources of uncertainties, such as input, parameter, and model structure (Schoups & Vrugt, 2010; Sorooshian et al., 1983). In addition, REMs correct biases in hydrological models (W. Li et al., 2017). The post-processing strategy differs from the classical joint inference approach, which estimates all parameters (hydrologic and error model) simultaneously using a single likelihood function (Bates & Campbell, 2001; Evin et al., 2014; Smith et al., 2015). The residual error model post-processor (REM-PP) is an attractive approach for converting deterministic into probabilistic predictions for two reasons. First, REM-PP avoids the need to re-calibrate the hydrological model. Second, REM-PP eliminates challenging interactions between water balance parameters and REM parameters (Evin et al., 2014).

Several studies have improved hydrological predictions using the REM strategy (Evin et al., 2014; Koutsoyiannis & Montanari, 2022; Kuczera et al., 2006; M. Li et al., 2016; McInerney et al., 2017; Reichert & Schuwirth, 2012; Todini, 2008; Wang et al., 2012; Zhao et al., 2015). In general, two approaches exist to describe residual model errors: (i) a lumped approach that aggregates all sources of uncertainty into an individual model error term (Montanari & Koutsoyiannis, 2012; Vrugt et al., 2022) and (ii) decomposition approaches, which attempt to disentangle the sources of uncertainty (Ajami et al., 2007; Kuczera et al., 2006). From a theoretical perspective, decomposition approaches are more versatile as they provide a comprehensive assessment of uncertainty estimation and identify specific sources of uncertainties. However, these approaches are complex to implement, demand more experience, and are typically more time-consuming in terms of both computational cost and user time. For these reasons, decomposition approaches tend not to be applied in practice. In contrast, lumped approaches evaluate the total uncertainty as estimated from the model-to-observation misfit. These approaches are more adequate for operational hydrology because they are less time-consuming, computationally efficient, and require little experience.

The post-processor method has been found to be robust in several previous studies (e.g., Evin et al., 2014; M. Li et al., 2016). On a daily scale, REMs showed a very good performance using power transformations (McInerney et al., 2017; Todini, 2008; Wang et al., 2012). In the post-processing strategy, the quality of uncertainty estimation has been shown (for single models) to depend on the residual error model and the objective function used to calibrate the hydrological model (Hunter et al., 2021), and characteristics of data (Jiang et al., 2019). The proportion of parametric uncertainty to total predictive uncertainty in streamflow is mostly modest (Kuczera et al., 2006; Sun et al., 2017; Yang et al., 2007) when parsimonious hydrological models are calibrated to long observed time series using a residual error model. In this context, predictions in operational hydrology habitually omit parameter uncertainty and focus on residual errors (Engeland & Steinsland, 2014; McInerney et al., 2017). This was the method employed in this research, where deterministic calibration was assumed solely via

optimization. In other words, the parameter estimation method was applied without trying to compute parametric uncertainty.

Large-sample hydrologic model intercomparison studies provide a unique opportunity to rank models, especially using standardized experiments. The Great Lakes region is a transboundary domain between the USA and Canada and has been the subject of several model intercomparison studies (Fry et al., 2014; Gaborit et al., 2017; Mai et al., 2021, 2022). Particularly, the Great Lakes Runoff Intercomparison Project phase 4: the Great Lakes (GRIP-GL) delivered a standardized experimental setup using similar geophysical datasets, forcings, a common routing product, and identical locations of performance assessment across the 10^6 km^2 study domain. This project evaluated 13 diverse streamflow models involving a large range of model types from machine-learning-based, lumped, subbasin-based, and gridded models that are either locally, regionally or globally calibrated (Mai et al., 2022). Like most past hydrological modelling intercomparison studies (e.g., Lake Michigan (GRIP-M; Fry et al., 2014), Lake Ontario (GRIP-O; Gaborit et al., 2017), and Lake Erie (GRIP-E; Mai et al., 2021)), GRIP-GL focused on deterministic model performance and did not consider the uncertainty of model streamflow predictions. Therefore, the GRIP-GL project dataset provides a unique chance to assess the relationship between model deterministic performance (assessed in GRIP-GL) and probabilistic performance (assessed here). This is the first study to the authors knowledge to apply uncertainty analysis to such a large variety of both models (13) and watersheds (141).

In contrast to GRIP-GL project, much of the research on hydrologic model calibration under uncertainty has been developed and assessed using only a single hydrological model, making it difficult to generalize. Certainly, the majority of previous studies was limited to using only one rainfall-runoff model (Evin et al., 2014; Jiang et al., 2019; D. Li et al., 2021; M. Li et al., 2016; McInerney et al., 2017, 2020, 2021; Renard et al., 2010; Salamon & Feyen, 2009; Thyer et al., 2009; Todini, 2008; Weerts et al., 2011; Woldemeskel et al., 2018). To our best knowledge, no studies have evaluated the impact of streamflow model choice on probabilistic predictions using multiple models and basins. This is the first essential research gap addressed in this investigation. Few studies have investigated the predictive uncertainty quantification using different hydrological models (e.g., Ye et al., (2014) and McInerney et al., (2017)). Ye et al., (2014) compare the effects of post-processing and model calibration on improving water forecasts under different hydroclimatic conditions and models . However, their study reduced comparisons to a deterministic perspective (comparing ensemble means for example). In addition, although they looked at results under multiple hydroclimatic conditions and seven hydrologic models, they did not evaluate the influence of the hydrologic model on the probabilistic prediction (e.g., no comparison of hydrologic model quality with and without post-processing). McInerney et al., (2017) and Morawietz et al., (2011) assessed several residual error models but used only two hydrological models, mainly HBV and GR4J. Neither study focused their analysis on a comparative assessment of relative model performance under deterministic versus probabilistic predictions, nor did they include machine learning models for streamflow like long short-term memory (LSTM) networks in their experimental design.

This studyexamines the effect of streamflow model choice on probabilistic predictions using a residual error model post-processing (REM-PP) approach to multiple streamflow models and basins. We addressed the following question: How does streamflow model choice impact probabilistic predictions? Finally, this study provides new insights into the influence of

hydrologic model choice on estimations of probabilistic streamflow predictions and changes in a model performance ranking.

2 Material and methods

We apply REM-PP to time series of streamflow in 141 calibration basins simulated using 13 diverse streamflow models. We adopt the Box-Cox (BC0.2) transformation residual error model as the REM-PP because it is an established practice in the hydrology literature (e.g., McInerney et al., (2017), McInerney et al., (2018), and Hunter et al., (2021)). McInerney et al., (2017) showed that this error model is effective for probabilistic predictions and uncertainty estimation in that it helps to remove heteroscedasticity and to deal with time series autocorrelation. In fact, McInerney et al., (2017) evaluated various error schemes, including standard and weighted least square, the Box-Cox transformations (with fixed and calibrated power parameters), and the log-sinh transformation across 23 basins. They recommended the Box-Cox approach due to it having the highest average performance and achieving high-quality daily probabilistic predictions. Figure 1 shows a graphical summary of the processes and methods followed in this study, and a detailed description of the REM-PP is provided in the following section.

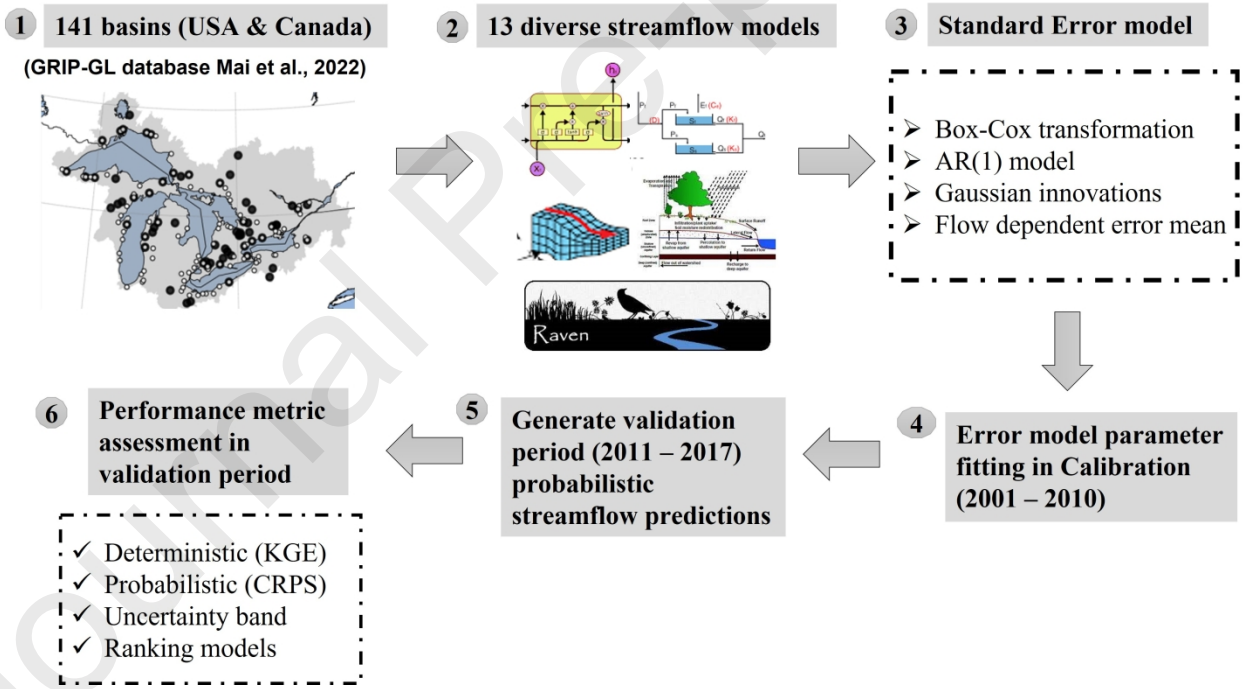


Figure 1. Flowchart for methods employed in this study.

2.1 Streamflow post-processing model

Consider a deterministic hydrological model (h) that represent streamflow predictions time series $\left(q_t^{\theta_h}\right)$with hydrological parameters (θ_h) and forcings (x).

$$q_t^{\theta_h} = h(\theta_h;x) \tag{1}$$

Then, define a probabilistic model of streamflow (Q_t), which is formulated from a deterministic hydrological model (h) by adding a random residual error term η_t intended to characterize the predictive uncertainty due to the combined effect of uncertainties in the input data, model parameters, and model structure.

$$Q_t(\theta,x) = f\left(q_t^{\theta_h},\eta_t^{\theta_\eta}\right) \tag{2}$$

where θ is comprised of hydrological model parameters (θ_h) and error model parameters (θ_η). The streamflow post-processing procedure used in this study consists of fitting a statistical model to the streamflow residual errors, defined by the differences between the observed and predicted daily streamflow time series. Residual errors of hydrological models are commonly heteroscedastic (i.e., have larger errors in large flows), asymmetrical, non-Gaussian, and persistent (i.e., often possess several days of consecutive errors with the same sign and similar magnitude) (Bates & Campbell, 2001; Smith et al., 2015; Sorooshian & Dracup, 1980). In many cases, residual error models perform well at the daily scale using the Box-Cox transformation (McInerney et al., 2017) and a first-order autoregressive AR(1) model (Evin et al., 2014). The Box-Cox transformation is intended to reduce the degree of heteroscedasticity and asymmetry in the error distribution, while the AR model is used to minimize the presence of persistence.

The probabilistic model Q_t is constructed by linking a deterministic component $q_t^{\theta_h}$ and a random residual error component η_t. Here, we formulate the residual error model as additive in transformed space,

$$z(Q_t;\theta_z) = z\left(q_t^{\theta_h};\theta_z\right) + \eta_t \tag{3}$$

where z is a transformation function with parameters θ_z. The deterministic discharge time series $q_t^{\theta_h}$ is generated by a hydrological model, which is a function of hydrological model parameters θ_h and forcings x_t. We use the Box-Cox transformation to reduce the heteroscedasticity and skewness in the residuals,

$$z(q_t;\theta_z) = \begin{cases} \dfrac{(q_t + A)^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \log(q_t + A), & otherwise \end{cases} \tag{4}$$

with parameters $\theta_z = \{A,\lambda\}$, where A is a shift or offset parameter and λ is a power parameter (Box & Cox, 1964). Following the recommendation of McInerney et al., (2017), the parameters

194 are fixed a priori at $\lambda = 0.2$ and $A = 0$. The temporal persistence of residual errors is modelled as
195 a first-order autoregressive (AR1) model,

$$\eta_t = \mu_t + \phi_\eta(\eta_{t-1} - \mu_{t-1}) + y_t \quad (5)$$

196 where ϕ_η is the lag-1 autoregressive parameter, μ_t is the mean of residuals at time t, and y_t
197 represents the innovation, which is a random component or noise term, at time t. The innovations
198 were assumed to follow a Gaussian distribution with a mean μ_t and standard deviation σ_y^2,

$$y_t \sim \mathcal{N}\left(\mu_t, \sigma_y^2\right) \quad (6)$$

199 In this study, we consider that the residual mean μ_t is flow-dependent. We assumed a
200 linear dependence with respect to the transformed streamflow (Jiang et al., 2019; M. Li et al.,
201 2017; Romero-Cuellar et al., 2019),

$$\mu_t = \alpha + \beta z\left(\theta_z, q_t^{\theta_h}\right) \quad (7)$$

202 where parameters α and β symbolize the intercept and slope respectively. The parameters of the
203 residual error model are indicated as $\theta_\eta = \{\theta_\mu, \sigma_y, \phi_\eta\}$, where θ_μ represents parameters of the
204 model for the mean μ_t. These parameters are grouped as $\theta_\mu = \{\alpha, \beta\}$. The full parameter set of
205 the probabilistic model Q_t is $\theta = \{\theta_h, \theta_z, \theta_\eta\}$. The REM-PP was implemented using the
206 "ProbPred" R package (Hunter et al., 2021).

207 2.1.1 Parameter estimation

208 The hydrologic model parameters θ were inferred from observed model forcings $\tilde{x} =$
209 $\{\tilde{x}_t; t = 1,\ldots,N_t\}$ and observed streamflow time series $\tilde{q} = \{\tilde{q}_t; t = 1,\ldots,N_t\}$, where N_t was the total
210 number of time steps. We used a two-stage post-processing approach for parameter estimation of
211 hydrological models and then the residual error model. In stage 1, the hydrological model
212 parameters θ_h were estimated previously by the modelling teams in GRIP-GL, typically by using
213 optimization algorithms (see Supplements of Mai et al., (2022) for full details for each model).
214 Some of the models were calibrated individually at each site, such that parameters were basin-
215 specific (local optimization). Other models were calibrated simultaneously across all sites (global
216 optimization). In stage 2, we estimated the residual error model parameters θ_η using the method-
217 of-moments (Hazelton, 2011) (see the appendix A of Hunter et al., (2021) for details on the
218 method-of-moments used in this study). Stage 2 computations are extremely fast because it
219 works only with observed data and optimal streamflow predictions from stage 1. Therefore, it
220 does not require additional hydrological model runs. The calibration period used was from
221 January 2001 to December 2010, while the validation period was from January 2011 to
222 December 2017.

223 2.1.2 Probabilistic predictions

224 The predictive distribution of streamflow is delivered by the probabilistic model Q_t
225 combining the deterministic hydrological model prediction and residual error probability model.
226 At time step t, assuming the deterministic streamflow prediction $q_t^{\widehat{\theta_h}}$ has been calculated, a
227 sample of size 1 from the predictive streamflow distribution, $q_t^{pred(s)}$, is produced as follows:

228 1) Sample random components (innovations) from a Gaussian distribution

$$y_t^{(s)} \leftarrow \mathcal{N}\left(\hat{\mu}_t, \hat{\sigma}_y^2\right) \tag{8}$$

229 2) Estimate residuals using Equation (5)

$$\eta_t^{(s)} = \hat{\mu}_t + \hat{\phi}_\eta\left(\eta_{t-1}^{(s)} - \hat{\mu}_{t-1}\right) + y_t^{(s)} \tag{9}$$

230 Note that for $t = 1$, we sample $\eta_1^{(s)} \leftarrow \mathcal{N}\left(0, \hat{\sigma}_y^2\right)$

231 3) Employ the inverse transformation

$$q_t^{pred(s)} = z^{-1}\left[z\left(q_t^{\widehat{\theta_h}}; \theta_z\right) + \eta_t^{(s)}; \theta_z\right] \tag{10}$$

232 The predictive distribution on day *t* can be characterized by repeating the above sampling
233 steps *s* times and then the complete sample of size *s* characterizing the predictive streamflow
234 distribution on day *t* is

$$q^{pred} = \left\{q_t^{pred(s)}; t = 1,...,T; s = 1,...,N_s\right\} \tag{11}$$

235 where N_s is the number of samples.

236 In general, this procedure describes a post-processor approach for predictive uncertainty
237 estimation. It provides practitioners with a straightforward method to convert deterministic
238 predictions to probabilistic ones using previously calibrated hydrological models (McInerney et
239 al., 2018).

240 2.2 Data, streamflow models, and study domain

241 Data used in this study were simulated and observed daily streamflows from the Great
242 Lakes Runoff Intercomparison Project phase 4: the Great Lakes (GRIP-GL) database (Mai et al.,

2022). The case study considers a subset of 141 calibration gauges of the GRIP-GL database. The study period consists of a warmup period (2000), a calibration period from January 2001 to December 2010, and a validation period from January 2011 to December 2017. Note that the number of gauges used in this study was different from the GRIP-GL database because we did not use spatial validation gauges. The GRIP-GL database contains 13 different streamflow models (Table 1).

The 13 streamflow models were grouped according to their main calibration strategy. The first group is the machine-learning-based model, which used a global calibration strategy. The global calibration means that the model was trained for all calibration stations at the same time, resulting in a single trained model for the entire study domain. The second group is comprised of the seven models that are locally calibrated, which means that the models were trained for each of the calibration stations individually. This leads to one calibrated model setup per gauging station basin. The third group is the five models that followed a regional calibration strategy, which means simultaneously calibrating model parameters to all calibration stations within a sub-region of the entire study region (rather than estimating parameters for each gauged basin individually).

The streamflow model parameters were generally estimated in Mai et al. (2022) (in stage 1) by optimization algorithms which minimize the aggregated differences between observed and simulated streamflow values. The optimization algorithms varied between models. Following this GRIP-GL split sample, all residual error models here are calibrated on the same calibration period and all results are analysed for the validation period. Detailed model descriptions can be found in Mai et al., (2022).

Table 1. Streamflow model structures investigated in the study. Models were set up, calibrated, and outputs derived in GRIP-GL (Mai et al., 2022).

Model ID	**Model name**	**Calibration Strategy**	**Model Type**	**Spatial resolution**
LSTM.lumped	Long Short-Term Memory network	Global	Machine learning	Basins
LBRM.CC.lumped	The Large Basin Runoff Model	Local	Conceptual model	Basins
HYMOD2.lumped	HYMOD2	Local	Conceptual model	Basins
GR4J.lumped	mode'le du GeÂ´nie Rural a' 4 parame'tres Journalier	Local	Conceptual model	Basins

HMETS.lumped	Hydrological Model of École de technologie supérieure	Local	Conceptual model	Basins
Blended.lumped	Blended aggregated	Local	Conceptual model with weights	Basins
Blended.Raven	Blended semi-distributed	Local	Conceptual model with weights	Subbasins
VIC.Raven	Variable Infiltration Capacity	Local	Water and energy balance	Grid
SWAT.Raven	The Soil and Water Assessment Tool	Regional	Physically based model	Subbasins
WATFLOOD.Raven	WATFLOOD	Regional	Physically based model	Grid
MESH.CLASS.Raven	Modélisation Environnementale communautaire-Surface and Hydrology	Regional	Conceptual Hydrology - Land Surface Model	Grid
MESH.SVS.Raven	Modélisation Environnementale communautaire-Surface and Hydrology	Regional	Conceptual Hydrology - Land Surface Model	Grid
GEM.Hydro.Watroute	GEM-Hydro	Regional	Physically based model	Grid

2.3 Performance metrics

Various deterministic and probabilistic performance metrics are evaluated for the validation period (1 January 2011 – 31 December 2017). Deterministic predictions were evaluated using the Kling-Gupta efficiency (KGE) (Gupta et al., 2009), while probabilistic predictions were evaluated using the mean continuous ranked probability score (CRPS) (Hersbach, 2000).

2.3.1 Kling-Gupta efficiency (KGE)

KGE was developed by decomposing the Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) into three parts, that is, correlation, mean bias, and variability bias:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (12)$$

where r symbolises the correlation between predicted and observed datasets, β represents the ratio of prediction mean to observation mean, and γ represents the ratio of standard deviation of predictions to observations. KGE is a positively oriented score with an ideal value of 1.

2.3.2 Continuous Ranked Probability Score (CRPS)

The CRPS is a familiar metric to assess the complete performance of probabilistic predictions. It measures the quadratic difference between the predicted and the ccumulative distribution function (CDF) of probabilistic predictions. The CRPS has the dimension of the predicted variable, i.e., CRPS' units are the same as area-normalized streamflow (mm d^{-1}). In addition, the CRPS can be decomposed into a reliability part and a sharpness part (Hersbach, 2000). For deterministic predictions, the CRPS is equivalent to the mean absolute error (MAE). In practice, the mean CRPS is defined as the average CRPS over the entire evaluation period because the CRPS measures the prediction for a single time step. We compute the CRPS from the empirical CDF of probabilistic predictions.

$$CRPS_i = \int_{-\infty}^{\infty} (F(q_t) - H_i(q_t \geq \tilde{q_t}))^2 dq_t \quad (13)$$

where $F(q_t)$ symbolises the cdf of the simulated flow predictive distribution at time t and H_i is the Heaviside step function, which equal 1 when the predictions are greater than the observations and equals 0 otherwise. q_t is the simulated flow and $\tilde{q_t}$ is the observed flow at time t. Note that we evaluate and report the mean $CRPS_i$ across all time steps in the validation period. In general, a high-quality prediction should have a CRPS as low as possible. The CRPS was calculated using the "scoringRules" R package (Jordan et al., 2019).

2.4 Evaluation of performance differences between models

Statistical significance testing was used to test if the streamflow models had better or worse performance metrics over the range of basins. Practical significance testing assesses whether the difference between performance metric values is large enough to represent substantial differences in a practical and necessarily subjective sense (Vaske et al., 2010). We used practical significance testing to evaluate differences in performance metrics. We decided that a difference by more than 10% of the change metric value for the hydrological models was practically significant. The practical significance testing is implemented using the method of McInerney et al., (2019). The pairwise Wilcoxon signed rank test (Bauer, 1972) was used to compare between distributions of a performance metric over all basins.

2.5 Normalised performance metrics

To facilitate a comparison between deterministic and probabilistic performance metrics in Section 3.3, KGE and CRPS were normalised for each basin according to this equation:

$$z_{KGE_i} = \frac{KGE_i - \min(KGE)}{\max(KGE) - \min(KGE)}; \quad (14)$$

$$z_{CRPS_i} = 1 - \left(\frac{CRPS_i - \min(CRPS)}{\max(CRPS) - \min(CRPS)}\right)$$

where KGE_i, $\min(KGE)$, and $\max(KGE)$ is the KGE metric value for model i, the minimum KGE across all models applied to that basin, and the maximum KGE across all models applied to that basin, respectively. Notation for CRPS is the same except note that CRPS refers to the average CRPS across all time steps. After we normalised the KGE and CRPS by basin, we compute the median for each model. Performance metrics range are $[0, 1]$, where 1 is the best and 0 is the worst.

3 Results

3.1. Representative example (single station)

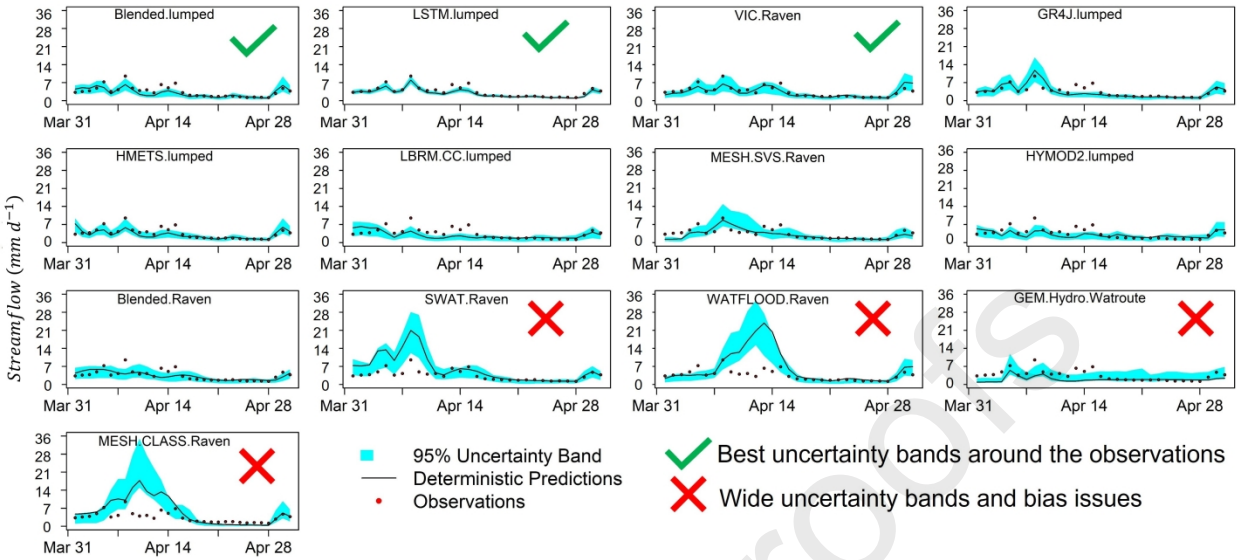
This section presents deterministic and probabilistic performance metrics of the 13 streamflow models for the Ganaraska River above Dale (Gauge 02HD012) during the validation period (1 January 2011 – 31 December 2017). Results indicate that the variation between model structures was wide in terms of deterministic and probabilistic performance metrics (Table 2). Specifically, KGE ranged from 0.816 (Blended.lumped) to 0.153 (MESH.CLASS.Raven) with 81% change and CRPS ranged from 0.180 (LSTM.lumped) to 0.299 (MESH.SVS.Raven) with 40% change. As seen in Table 2, deterministic model skill (KGE) is correlated to probabilistic prediction skill (CRPS). In the representative station (02HD012) where the deterministic model performs well, the performance of the probabilistic predictions in terms of the CRPS will generally be good (e.g., LSTM.lumped). In contrast, when the deterministic model performs poorly, the flow uncertainty estimation is also poor (e.g., MESH.CLASS.Raven). This result shows that the residual error model post-processor could not "remedy" a poor deterministic model prediction.

Table 2. Deterministic and probabilistic performance metrics of streamflow model predictions based on 13 model structures for the single station (02HD012) during validation period (1 January 2011 – 31 December 2017).

Model	**Metric**	
	KGE	**CRPS**

Blended.lumped	0.816	0.216
LSTM.lumped	0.803	0.180
VIC.Raven	0.802	0.240
GR4J.lumped	0.802	0.223
HMETS.lumped	0.774	0.234
LBRM.CC.lumped	0.730	0.280
MESH.SVS.Raven	0.667	0.299
HYMOD2.lumped	0.664	0.277
Blended.Raven	0.597	0.284
SWAT.Raven	0.241	0.265
WATFLOOD.Raven	0.175	0.270
GEM.Hydro.Watroute	0.168	0.289
MESH.CLASS.Raven	0.153	0.260

Figure 2 illustrates sample flow time series with 95% uncertainty band for 13 streamflow models during a freshet event in 2014. In this paper, 95% uncertainty bands were calculated to be within 2.5 and 97.5 percentiles, as calculated from the Gaussian error model in transform space. Figure 2 shows that the Blended.lumped, LSTM.lumped, and VIC.Raven models produce the best predictions, with a narrow 95% uncertainty band around the observed data. In contrast, MESH.CLASS.Raven, SWAT.Raven, and WATFLOOD.Raven models present wide 95% uncertainty bands. These models are over-confident, with 44%, 41%, and 29%, respectively, of observed flows outside the 95% uncertainty band. These models also show a bias issue, which overestimate the high flow and underestimate the low flow.



341

342 **Figure 2**. Sample time series with coloured uncertainty bands of streamflow based on 13 model
343 structures for the single station (02HD012) during validation period (1 January 2011 – 31
344 December 2017).

345 These results show the influence of streamflow model structures on streamflow
346 predictive uncertainty quantification (Figure 2). For the same station and using the same residual
347 error model post-processor, streamflow uncertainty quantification results were dissimilar for
348 different streamflow models.

349 3.2 Model performance evaluation using full set of basins

350 3.2.1 Deterministic streamflow predictions (KGE)

351 Figure 3 compares the KGE of 13 streamflow models across all 141 case study basins.
352 The LSTM.lumped model delivers the best KGE, indicated by the high median KGE
353 performance metric of 0.8. The LSTM.lumped model also provides practically and statistically
354 significant improvements in KGE. This outcome is consistent with the results of Mai et al.,
355 (2022), Kratzert et al., (2018), and Kratzert et al., (2019). These studies have shown that
356 globally-optimized LSTM models generally outperform conventional physically-based
357 hydrological models used in water resource modelling. The Blended.lumped, Blended.raven, and
358 HYMOD2.lumped models achieve a similar performance with median values of KGE 0.72, 0.72,
359 and 0.71 respectively. In contrast to the LSTM.lumped model, the MESH.CLASS.Raven model
360 provides the lowest KGE with median value of 0.42. It is interesting that the newer structurally-
361 flexible models such as LSTM and Blended provide top performance in terms of the
362 deterministic performance metric (KGE).

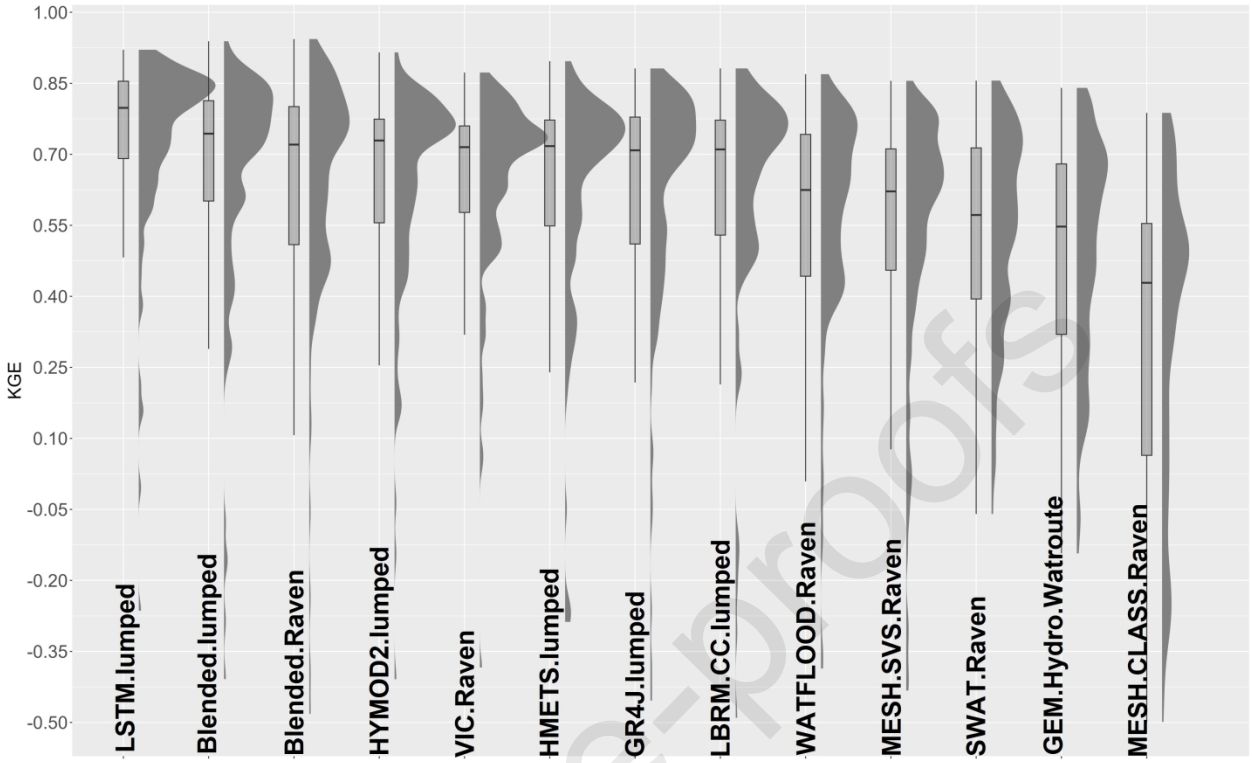


Figure 3. Density curves and boxplots of mean KGE over 13 streamflow model structures for all case study basins during validation period (1 January 2011 – 31 December 2017). For each model, the distribution of mean KGE scores across basins is summarized by a box and whisker plot highlighting the median and the 25 and 75 quantiles as well as the entire empirical distribution.

3.2.2 Probabilistic streamflow predictions (CRPS)

Figure 4 reports the CRPS of 13 streamflow models overall case study basins. The LSTM.lumped model is, once again, the best-performing model, with a median CRPS metric of 0.25, which is practically and statistically significant better than the other numerical models. On the contrary, the MESH.CLASS.Raven model performed generally worse, with a median CRPS metric of 0.46.

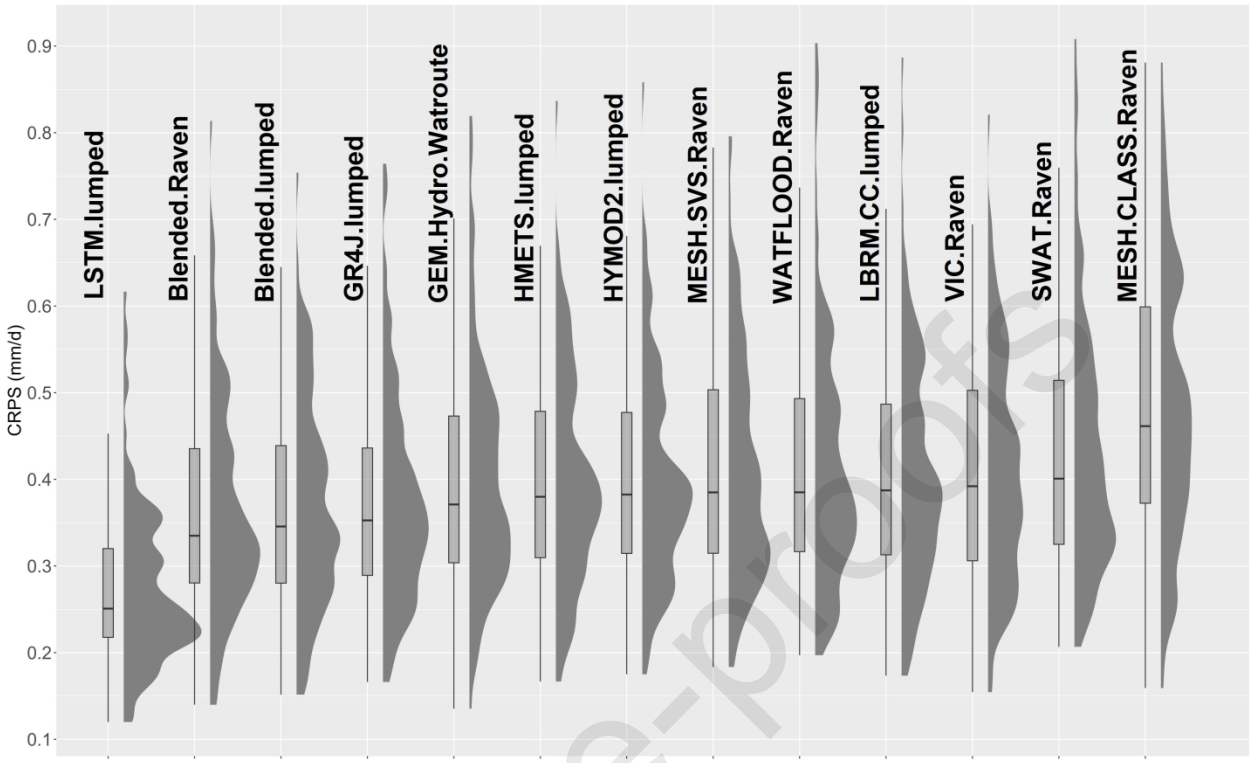


Figure 4. Density curves and boxplots of mean CRPS over 13 streamflow model structures for all case study basins during validation period (1 January 2011 – 31 December 2017). For each model, the distribution of mean CRPS scores across basins is summarized by a box and whisker plot highlighting the median and the 25 and 75 quantiles as well as the entire empirical distribution.

Figure 5 shows a heatmap of KGE (top) and CRPS (bottom) across the basins and models during the validation period. This figure helps to explore the distributions of figure 3 and 4 in detail at the individual basin level. Results confirm that LSTM.lumped delivers better KGE and CRPS values in many basins (blue tones), while MESH.CLASS.Raven provides the lowest KGE and CRPS values in many basins (red tones). Results also suggest that after post-processing there is less variation by models because there are vertical strips of the same color on the bottom (after post-processing) versus the colour variations on the top (no post-processing). This pattern seems especially true for basins with poor KGE (basins located in the right part of figure 5) and for basins with diverse KGE (e.g., 02HC025, 02AB006, and 02GA047 which are highlighted on the x-axis with black squares). A possible explanation for this pattern might be the rationality behind the REM-PP. In most cases, REM-PP struggles to match model outputs (e.g., streamflow) with observations. These matches employ statistical models that reduce model biases from all sources of uncertainty and characterize the associated uncertainty (Ye et al., 2014). Therefore, if REM-PP works reasonably well, it is sensible that REM-PP's outputs show similar performance. In other words, for some basins, the post-processing can equalise all the model's performance in terms of CRPS. This explanation is in line with the work of Ehlers et al., (2019) who found that the k-NN post-processor was able to compensate a poor deterministic simulation. Results also

indicate that in some basins it is difficult to obtain satisfactory performance in terms of KGE (dark grey cells) due to these basins being highly flow regulated. Similarly, in terms of CRPS, some basins seem difficult to obtain satisfactory performance; these basins show spiky hydrographs (e.g., 02FE009 and 02LB032 and these are highlighted on the x-axis with red squares).

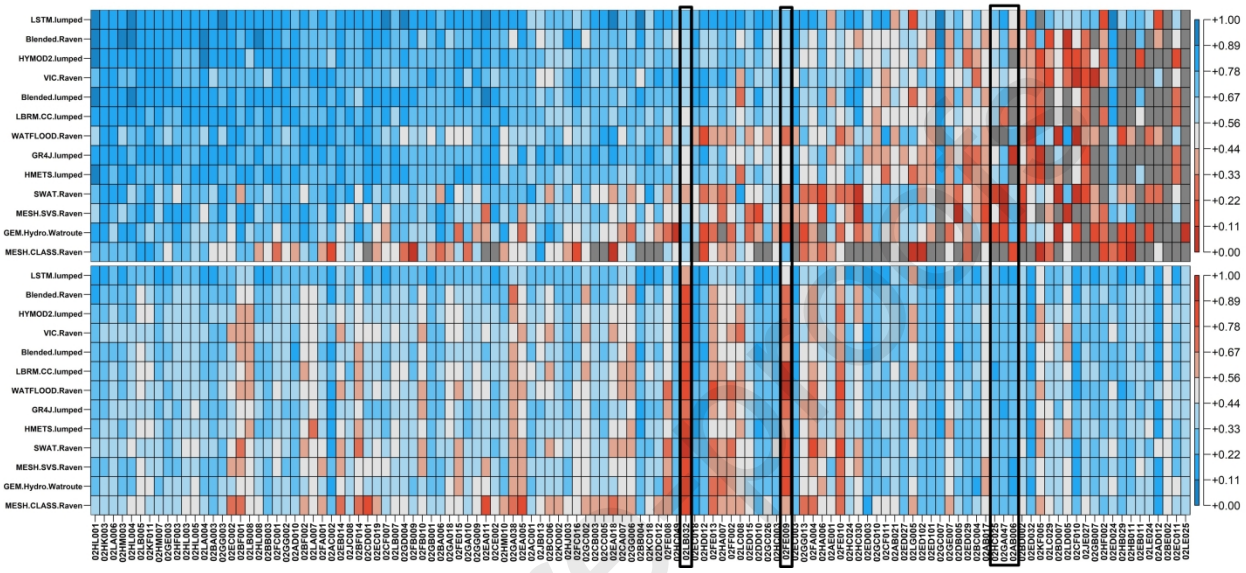


Figure 5. Heatmap of KGE (top) and CRPS (bottom) over 13 streamflow model structures for all case study basins during the validation period (1 January 2011 – 31 December 2017).Horizontal axes denote different basins, which are ordered according to the mean KGE performance metric (best on the left), while vertical axes denote different models. Dark grey cells symbolize values of KGE below 0 meaning undesirable values. Blue tones indicate good (preferred) scores, and red tones signify bad scores. Some basin IDs are highlighted with black squares because they present particular behavior, which is examined in the result section.

3.3 Ranking models

Figure 6 represents the normalised performance metric of the KGE and CRPS of 13 streamflow models across all case study basins during the validation period (1 January 2011 – 31 December 2017). Results show that LSTM.lumped model provides the best performance across deterministic (KGE) and probabilistic (CRPS) metrics. In contrast to the LSTM.lumped model, the MESH.CLASS.Raven model delivers the poorest performance in terms of KGE and CRPS. The LSTM.lumped model is consistently the best-performing model. In addition, Figure 6 shows the ranking based on the median normalised KGE and CRPS. The order of the numerical models was determined based on the ranking of the median normalised KGE.

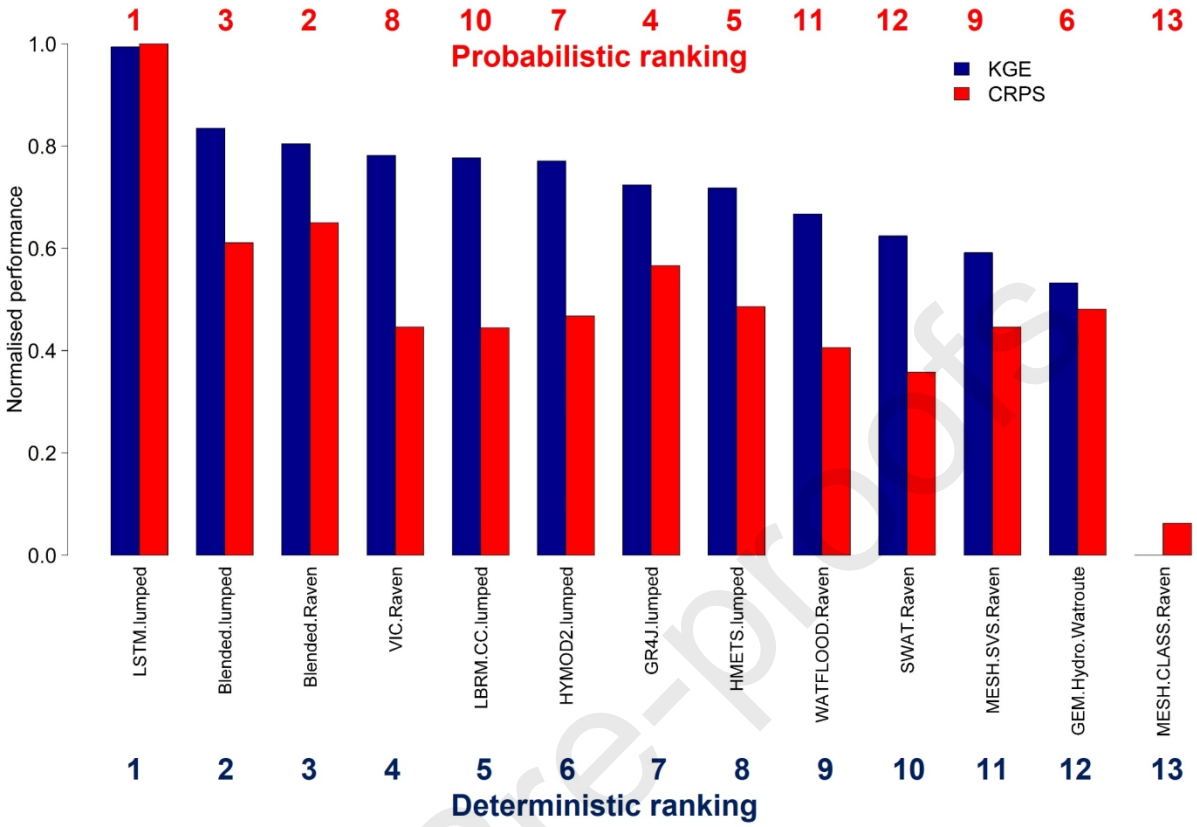


Figure 6. Comparing the model ranking and normalised deterministic and probabilistic performance of 13 streamflow model structures during the validation period (1 January 2011 – 31 December 2017) using the median of the normalised KGE and CRPS for all case study. Red numbers in the top represent the probabilistic ranking (CRPS) and blue numbers in the bottom represent the deterministic ranking (KGE). Smaller number in the ranking indicates better performance.

A key finding is that moving from deterministic to probabilistic predictions using a post-processor strategy does not change the streamflow model performance ranking for the best and worst models in GRIP-GL, but it can change the ranks considerably for models of intermediate rank. In particular, post-processing greatly improves GR4J and GEM.Hydro.Watroute performance rank (relative to deterministic) while post-processing seems to degrade VIC and LBRM.CC.lumped rankings (relative to deterministic) (Figure 6). In part, these ranking changes reflect the comparable performance of probabilistic models in terms of CRPS – many of the scores fall within the range 0.4 to 0.5, where a swap in rank may simply be an artefact of the subtle variations in scores. However, results show a general pattern. When ranked by both KGE and CRPS, the highest ranked models are LSTM and Blended while MESH.CLASS, SWAT, and WATFLOOD are the lowest-ranked models (Figure 6). While not absolutely true, the probabilistic prediction quality tends to improve with deterministic performance, with both higher sharpness and robustness. As might be expected, poorly performing deterministic models tend to have much wider error models to compensate for their poorer performance, and therefore probabilistic prediction ranges are much wider. Inconsistency or performance likely leads to

reduced robustness of the models, as it becomes more difficult for a single error model to capture model uncertainty which varies seasonally or with varying flow regime.

These results illustrate that streamflow model choice strongly influences probabilistic performance. Different streamflow model structures produce varied deterministic performance metrics (KGE), and these deterministic performance differences produce distinct probabilistic performance metrics (CRPS) using the same residual error model post-processor. This variance of CRPS values indicates that the quality of probabilistic predictions depends strongly on the quality of the deterministic prediction. This finding is in line with Morawietz et al., (2011), who evaluated different versions of the autoregressive error model as post-processors for a probabilistic streamflow forecast system in Norway.

4 Discussion and conclusions

This research examines the effects of streamflow model choice on probabilistic performance (CRPS) using a residual error model post-processing (REM-PP) approach to multiple streamflow models and basins. We assess 141 calibration basins from the Great Lakes region in the USA and Canada. This assessment employed 13 diverse streamflow predictions based on machine-learning, lumped, subbasin, and gridded models. Using daily-scale streamflow model predictions, we apply an additive REM-PP in transformed space using the Box-Cox transformation, a first-order autoregressive AR(1) model, Gaussian innovations, and a flow-dependent error mean. The main findings are as follows:

(1) Streamflow model choice strongly influences the robustness of probabilistic predictions in terms of CRPS.

(2) Moving from deterministic to probabilistic predictions using a post-processor approach does not change the streamflow model performance ranking for the best and worst deterministic models, but models of intermediate rank in deterministic evaluation do not have consistent ranking when evaluated in probabilistic mode (Figure 6). This provides some evidence that probabilistic forecasts can, in some sense, compensate for differences in model performance.

(3) Post-processing residual errors of long short-term memory (LSTM) network models are consistently the best-performing model in terms of KGE (Figure 3) and CRPS (Figure 4, Figure 5, and Figure 6). Notably, the normalized performance gap between the LSTM and the second-best model is roughly doubled when the model comparison moves from a deterministic to a probabilistic perspective. That is, the LSTM model looks even more dominant over the other models when we use a post-processing approach to consider model prediction uncertainty (Figure 6).

Based on empirical case studies, these findings suggest that merging deterministic predictions of LSTM with the statistical modelling of the residual errors (post-processing) provides the best approach to convert deterministic predictions to probabilistic ones. This approach would be a promising route to characterize and reduce uncertainty in hydrological predictions. These results are in line with the benchmarking study of Mai et al., (2022) who showed that LSTM models outperformed the conventional hydrological models used to predict

daily streamflow. The superior performance of LSTM can be attributed to its capacity to digest more training data (flow locations, basin characteristics, and forcings) and its global setup. Note that LSTM was the only globally calibrated model. In addition, the range of error model parameters are similar across all streamflow models except for the LSTM, which appear to be unique. Boxplots of error model parameter values are showed in Appendix A.

In a general sense, model intercomparisons are sensitive to the performance evaluation framework. If the models are evaluated from a deterministic framework, the same performance of the models cannot be expected from a probabilistic framework. In other words, assuming that the deterministic model ranking will be the same when moving to a probabilistic performance evaluation is an inappropriate assumption. When probabilistic performance is implemented, models are ranked differently. The residual error model could be an integral component in model intercomparison. In past model intercomparisons, a few studies focussed on deterministic model performance ranking (e.g., Lake Michigan (GRIP-M; Fry et al., 2014), Lake Ontario (GRIP-O; Gaborit et al., 2017), and Lake Erie (GRIP-E; Mai et al., 2021)), but no previous study has investigated the change between deterministic and probabilistic model performance rankings. The novelty of this study lies in assessing the influence of multiple and diverse streamflow predictions (i.e., multiple hydrologic models) on the quality of uncertainty estimates relative to deterministic model performance. In other words, this study shows that streamflow model choice affects the uncertainty estimation using large-sample catchments and a standardized experimental setup.

In this work, we implemented a straightforward method to transform deterministic into probabilistic predictions. The method is based on the post-processor strategy for the residual error model, which is empirically more flexible and robust than joint strategies because it avoids challenging interactions between water balance parameters and error model parameters (Evin et al., 2014). The REM-PP used the Method-of-Moments inference method, which does not need a likelihood function or optimisation, and therefore demands less computational costs. Although the Method-of-Moments is simple, it can deliver predictive inference like Maximum Likelihood or Bayesian methods (McInerney et al., 2018). The REM-PP quantifies the predictive uncertainty without the need to change the objective function or to re-calibrate the hydrological model (Hunter et al., 2021).

This study has potential limitations. We only addressed one simple post-processing algorithm configuration. The results may not directly extend to other methods of uncertainty analysis, such as Monte Carlo, informal or formal Bayesian methods, etc. The results may not extend to different regions where models may rank differently in terms of accuracy.

Given that this research has focused on the effects of the hydrological model choice on streamflow predictive uncertainty quantification, future work should focus on a more comprehensive assessment using several hydroclimatological regimes. Especially, for example, is whether predictive uncertainty estimation depends more on the hydrological model or on the hydroclimatic regime.

Mainly in future work, we will include the uncertainty of model inputs (precipitation and temperature) using a stochastic modelling generator. Further development is needed to improve the predictive uncertainty estimation of low flows, including the treatment of zero flows

(McInerney et al., 2019). They can occur, for instance, in the extremely cold winter low flow conditions in the Arctic and Prairie basins, in Canada. Remarkably, future research also should focus on predictive uncertainty quantification at ungauged catchments which was not considered here.

In general, this study helps to provide insight into the influence of streamflow model choice on the estimation of prediction uncertainty and connects the gap between deterministic and probabilistic prediction methods in practical hydrological applications. Moreover, this research uses several case studies and streamflow models, including machine-learning-based, lumped, subbasin-based, and gridded models, and applies a straightforward method to researchers and practitioners who need to transform deterministic predictions to probabilistic ones to characterize predictive uncertainty and support the management of hydrological risk.

Finally, this study is the first to perform a large-sample model intercomparison (13 models and 141 basins) from both a deterministic and probabilistic perspective and then compare the differences in model rankings. The results of this study emphasise how crucial it is to combine residual error models with the best deterministic streamflow model predictions (like LSTM) to improve the quality and value of hydrological predictions, quantify uncertainty, and better inform decision-making in operational water management.

5 Data available

The streamflow data used for this analysis are available at https://hydrohub.org/grip-gl/maps_streamflow.html (accessed on July 10, 2023). The code to implement the residual error model is available via the "ProbPred" R package (https://github.com/AdWater/ProbApp).

CRediT authorship contribution statement

Jonathan Romero-Cuellar: Conceptualisation, Investigation, Formal analysis,Visualisation, Software, Methodology, Writing - original draft. **Rezgar Arabzadeh:** Software, Data curation, Visualisation, Formal analysis. **James R. Craig:** Conceptualisation, Methodology, Writing - review & editing, Supervision, Resources. **Bryan A. Tolson:** Conceptualisation, Methodology, Writing - review & editing, Supervision. **Juliane Mai:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is funded by the Natural Resources Canada Emergency Management program, the NSERC Canada Research Chairs program, and NSERC Discovery Individual grants held by Drs. Craig and Tolson. Special thanks to David McInerney for his support in ProbPred R package. We sincerely thank the associate editor and two anonymous reviewers for their thoughtful and helpful feedback, which significantly improved the manuscript.

Appendix A: REMs' parameter ranges

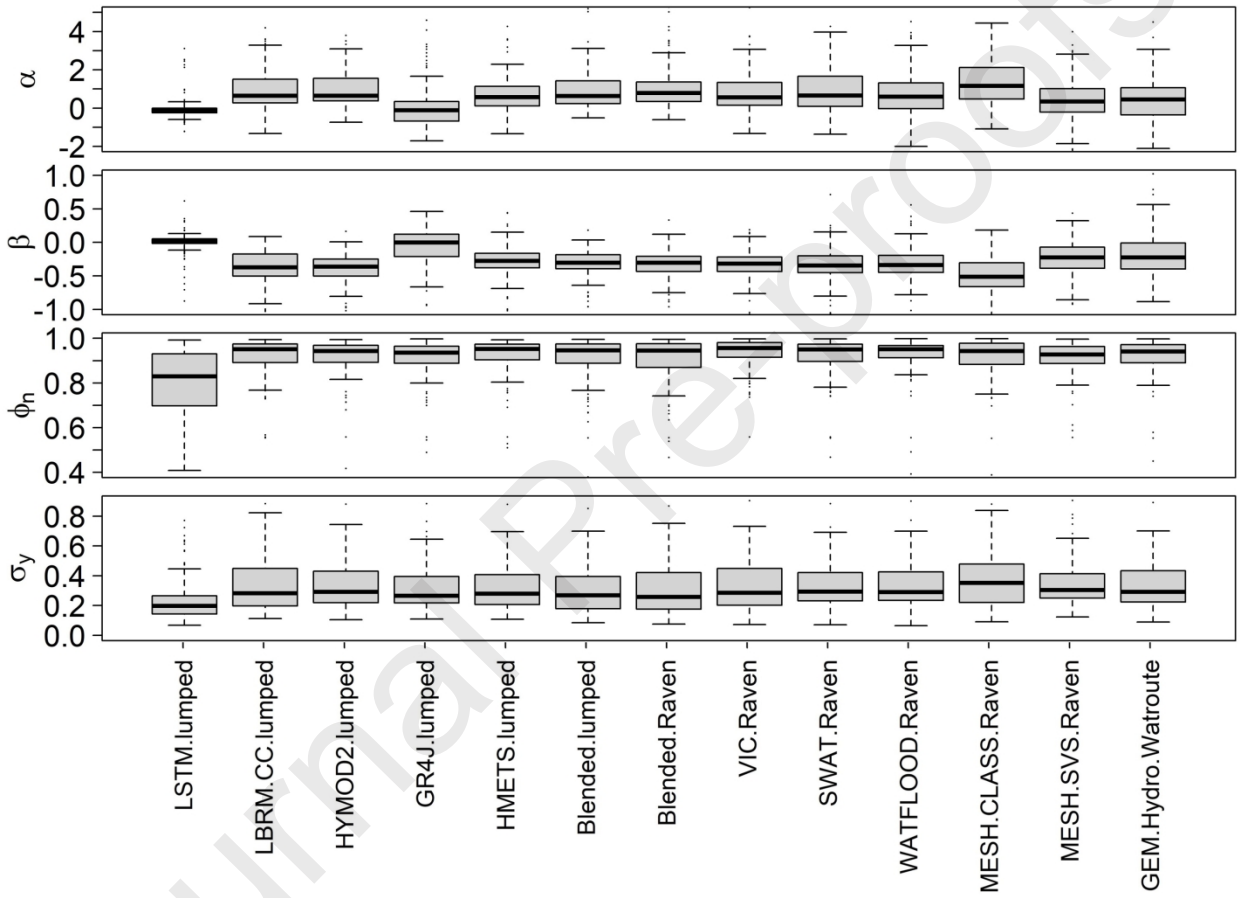


Figure A. Boxplots of error model parameter values across streamflow models. α is the intercept and β is the slope of flow linear regression function respectively. ϕ_η is the lag-1 autoregressive parameter, and σ_y standard deviation of Gaussian distribution.

References

Ajami, N. K., Duan, Q., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1), W01403. https://doi.org/10.1029/2005WR004745

Bastola, S., Murphy, C., & Fealy, R. (2012). Generating probabilistic estimates of hydrological response for Irish catchments using a weather generator and probabilistic climate change scenarios. *Hydrological Processes*, *26*(15), 2307–2321. https://doi.org/10.1002/hyp.8349

Bates, B. C., & Campbell, E. P. (2001). A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resources Research*, *37*(4), 937–947. https://doi.org/10.1029/2000WR900363

Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, *67*(339), 690. https://doi.org/10.2307/2284469

Box, G.E.P., & Cox, D.R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society*, Series B, 26 (2), 211–252. http://www.jstor.org/stable/2984418

Dias, L. F., Aparício, B. A., Nunes, J. P., Morais, I., Fonseca, A. L., Pastor, A. V., & Santos, F. D. (2020). Integrating a hydrological model into regional water policies: Co-creation of climate change dynamic adaptive policy pathways for water resources in southern Portugal. *Environmental Science and Policy*, *114*, 519–532. https://doi.org/10.1016/j.envsci.2020.09.020

Ehlers, L.B., Wani, O., Koch, J., Sonnenborg, T.O., & Refsgaard, J.C. (2019). Using a simple post-processor to predict residual uncertainty for multiple hydrological model outputs. *Advances in Water Resources*, 129, 16-30. https://doi.org/10.1016/j.advwatres.2019.05.003

Engeland, K., & Steinsland, I. (2014). Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resources Research*, *50*(1), 182–197. https://doi.org/10.1002/2012WR012757

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*, *50*(3), 2350–2375. https://doi.org/10.1002/2013WR014185

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*(7), 5619–5633. https://doi.org/10.1002/2016WR019129

Fry, L. M., Gronewold, A. D., Fortin, V., Buan, S., Clites, A. H., Luukkonen, C., Holtschlag, D., Diamond, L., Hunter, T., Seglenieks, F., Durnford, D., Dimitrijevic, M., Subich, C., Klyszejko, E., Kea, K., & Restrepo, P. (2014). The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M). *Journal of Hydrology*, *519*, 3448–3465. https://doi.org/10.1016/J.JHYDROL.2014.07.021

Gaborit, É., Fortin, V., Tolson, B., Fry, L., Hunter, T., & Gronewold, A. D. (2017). Great Lakes Runoff Inter-comparison Project, phase 2: Lake Ontario (GRIP-O). *Journal of Great Lakes Research*, *43*(2), 217–227. https://doi.org/10.1016/J.JGLR.2016.10.004

Gupta, A., & Govindaraju, R. S. (2023). Uncertainty quantification in watershed hydrology: Which method to use? *Journal of Hydrology*, 616, 128749. https://doi.org/10.1016/j.jhydrol.2022.128749

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hazelton, M. L. (2011). Methods of Moments Estimation BT. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 816–817). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

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*, *603*, 126578. https://doi.org/10.1016/J.JHYDROL.2021.126578

Jiang, X., Gupta, H. V., Liang, Z., & Li, B. (2019). Toward Improved Probabilistic Predictions for Flood Forecasts Generated Using Deterministic Models. *Water Resources Research*, *55*(11), 9519–9543. https://doi.org/10.1029/2019WR025477

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, *90*, 1–37. https://doi.org/10.18637/JSS.V090.I12

Koutsoyiannis, D., & Montanari, A. (2022). Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions. *Water Resources Research*, *58*(1), e2021WR031215. https://doi.org/10.1029/2021WR031215

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci*, *22*, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1–2), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010

Li, D., Marshall, L., Liang, Z., Sharma, A., & Zhou, Y. (2021). Characterizing distributed hydrological model residual errors using a probabilistic long short-term memory network. *Journal of Hydrology*, *603*. https://doi.org/10.1016/j.jhydrol.2021.126888

Li, M., Wang, Q. J., Bennett, J. C., & Robertson, D. E. (2016). Error reduction and representation in stages (ERRIS) in hydrological modelling for ensemble streamflow forecasting. *Hydrology and Earth System Sciences*, *20*(9), 3561–3579. https://doi.org/10.5194/hess-20-3561-2016

Li, M., Wang, Q. J., Robertson, D. E., & Bennett, J. C. (2017). Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs. *Journal of Hydrology*, *555*, 586–599. https://doi.org/10.1016/j.jhydrol.2017.10.057

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, *4*(6), e1246. https://doi.org/10.1002/wat2.1246

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., & Waddell, J. W. (2022). The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL). *Hydrology and Earth System Sciences*, *26*(13), 3537–3572. https://doi.org/10.5194/HESS-26-3537-2022

Mai, J., Tolson, B. A., Shen, H., Gaborit, É., Fortin, V., Gasset, N., Awoye, H., Stadnyk, T. A., Fry, L. M., Bradley, E. A., Seglenieks, F., Temgoua, A. G. T., Princz, D. G., Gharari, S., Haghnegahdar, A., Elshamy, M. E., Razavi, S., Gauch, M., Lin, J., … Pietroniro, A. (2021). Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E). *Journal of Hydrologic Engineering*, *26*(9), 05021020. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002097

McInerney, D., Kavetski, D., Thyer, M., Lerat, J., & Kuczera, G. (2019). Benefits of Explicit Treatment of Zero Flows in Probabilistic Hydrological Modeling of Ephemeral Catchments. *Water Resources Research*, *55*(12), 11035–11060. https://doi.org/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*, *109*, 306–314. https://doi.org/10.1016/J.ENVSOFT.2018.07.001

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. *Water Resources Research*, *56*, e2019WR026979. https://doi.org/10.1029/2019WR026979

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., & Kuczera, G. (2021). Improving the Reliability of Sub-Seasonal Forecasts of High and Low Flows by Using a Flow-Dependent Nonparametric Model. *Water Resources Research*, *57*, e2020WR029317. https://doi.org/10.1029/2020WR029317

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*, *53*(3), 2199–2239. https://doi.org/10.1002/2016WR019168

Moges, E., Demissie, Y., Larsen, L., & Yassin, F. (2021) Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis. *Water*, 13(1),28. https://doi.org/10.3390/w13010028

Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*, W09555. https://doi.org/10.1029/2011WR011412

Morawietz, M., Xu, C. Y., Gottschalk, L., & Tallaksen, L. M. (2011). Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system. *Journal of Hydrology*, *407*(1–4), 58–72. https://doi.org/10.1016/J.JHYDROL.2011.07.007

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Reichert, P., & Schuwirth, N. (2012). Linking statistical bias description to multiobjective model calibration. *Water Resources Research*, *48*(9), 9543. https://doi.org/10.1029/2011WR011391

Reggiani, P., Talbi, A., & Todini, E. (2022). Towards informed water resources planning and management. Hydrology, 9(8), 136. https://doi.org/10.3390/hydrology9080136

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5), W05521. https://doi.org/10.1029/2009WR008328

Romero-Cuellar, J., Abbruzzo, A., Adelfio, G., & Francés, F. (2019). Hydrological post-processing based on approximate Bayesian computation (ABC). *Stoch Environ Res Risk Assess*, *33*(7), 1361–1373. https://doi.org/10.1007/s00477-019-01694-y

Salamon, P., & Feyen, L. (2009). Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. *Journal of Hydrology*, *376*(3–4), 428–442. https://doi.org/10.1016/j.jhydrol.2009.07.051

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), W10531. https://doi.org/10.1029/2009WR008933

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation. *Water Resources Research*, *59*(2), e2022WR032201. https://doi.org/10.1029/2022WR032201

Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, *528*, 29–37. https://doi.org/10.1016/j.jhydrol.2015.05.051

Sorooshian, S., & Dracup, J. A. (1980). Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resources Research*, *16*(2), 430–442. http://doi.wiley.com/10.1029/WR016i002p00430

Sorooshian, S., Gupta, V. K., & Fulton, J. L. (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*, *19*(1), 251–259. https://doi.org/10.1029/WR019i001p00251

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*, *554*, 680–692. https://doi.org/10.1016/J.JHYDROL.2017.09.041

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2008WR006825

Todini, E. (2008). A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, *6*(2), 123–137. https://doi.org/10.1080/15715124.2008.9635342

Vaske, J. J., Gliner, J. A., & Morgan, G. A. (2010). Communicating Judgments About Practical Significance: Effect Size, Confidence Intervals and Odds Ratios. *Human Dimensions of Wildlife*, *7*(4), 287–300. https://doi.org/10.1080/10871200214752

Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. Water Security, 1, 28-35. https://doi.org/10.1016/j.wasec.2017.06.001

Vrugt, J. A., de Oliveira, D. Y., Schoups, G., & Diks, C. G. H. (2022). On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. *Journal of Hydrology*, *615*, 128542. https://doi.org/10.1016/J.JHYDROL.2022.128542

Wang, Q. J., Shrestha, D. L., Robertson, D. E., & Pokhrel, P. (2012). A log-sinh transformation for data normalization and variance stabilization. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR010973

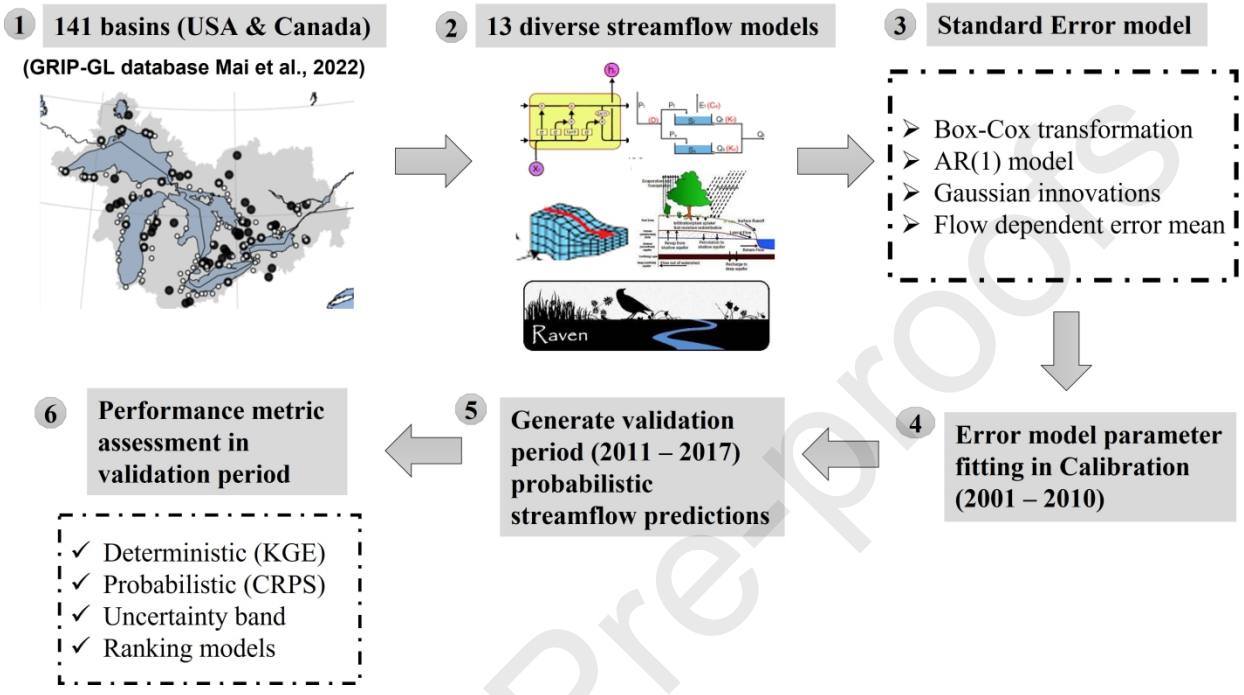
Weerts, A. H., Winsemius, H. C., & Verkade, J. S. (2011). Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrol. Earth Syst. Sci*, *15*, 255–265. https://doi.org/10.5194/hess-15-255-2011

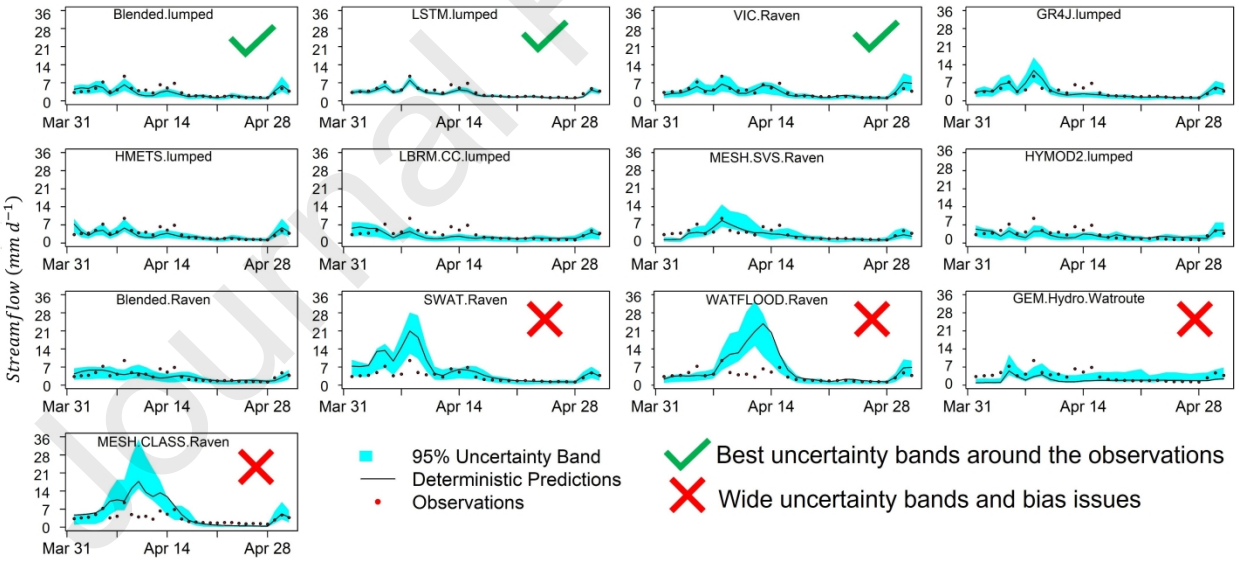
Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N., & Kuczera, G. (2018). Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *22*(12), 6257–6278. https://doi.org/10.5194/HESS-22-6257-2018

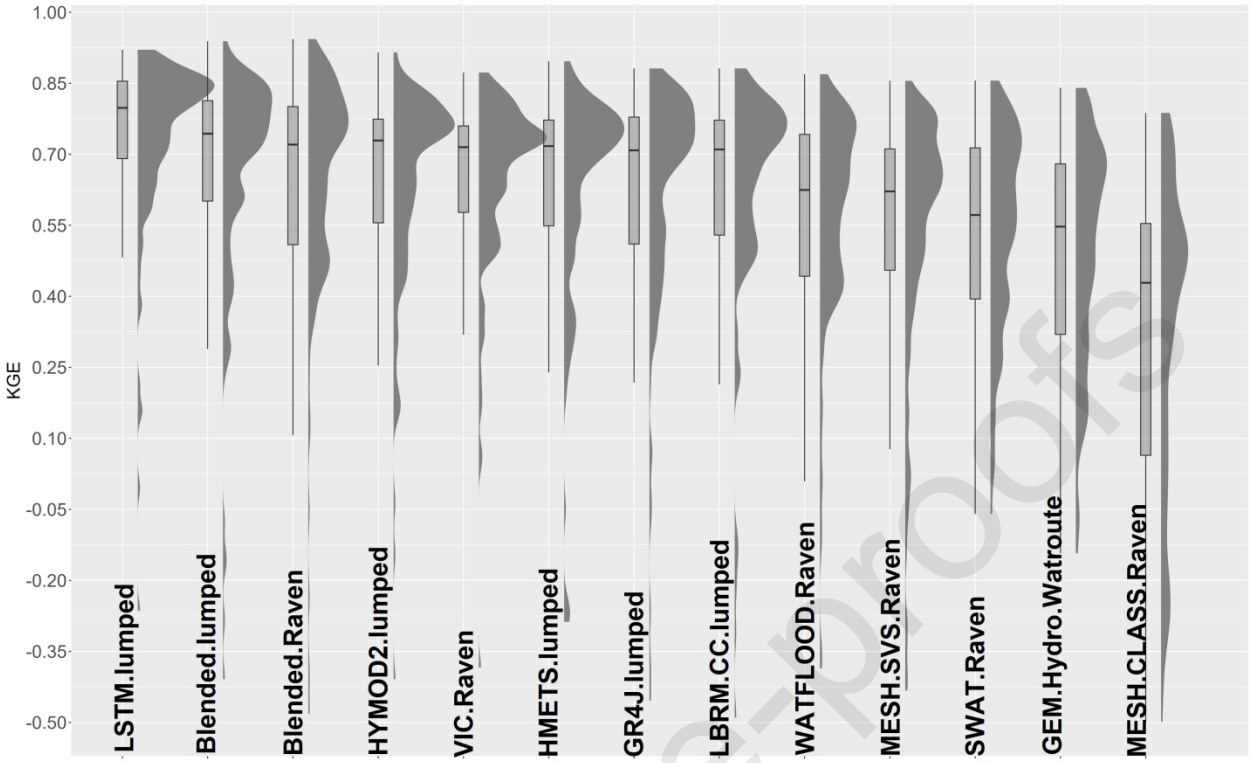
Yang, J., Reichert, P., & Abbaspour, K. C. (2007). Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland). *Water Resources Research*, *43*(10), W10401. https://doi.org/10.1029/2006WR005497

Ye, A., Duan, Q., Yuan, X., Wood, E. F., & Schaake, J. (2014). Hydrologic post-processing of MOPEX streamflow simulations. *Journal of Hydrology*, *508*, 147–156. https://doi.org/10.1016/j.jhydrol.2013.10.055

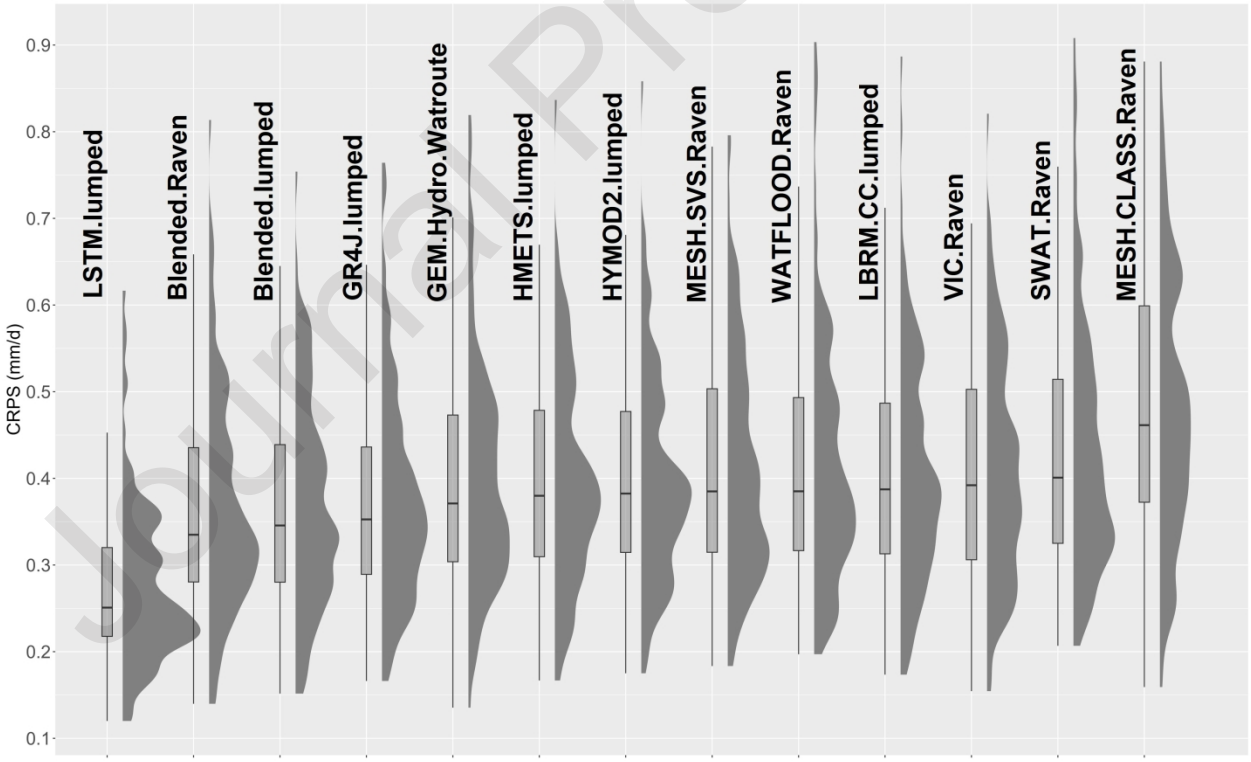
Zhao, T., Wang, Q. J., Bennett, J. C., Robertson, D. E., Shao, Q., & Zhao, J. (2015). Quantifying predictive uncertainty of streamflow forecasts based on a Bayesian joint probability model. *Journal of Hydrology*, *528*, 329–340. https://doi.org/10.1016/j.jhydrol.2015.06.043



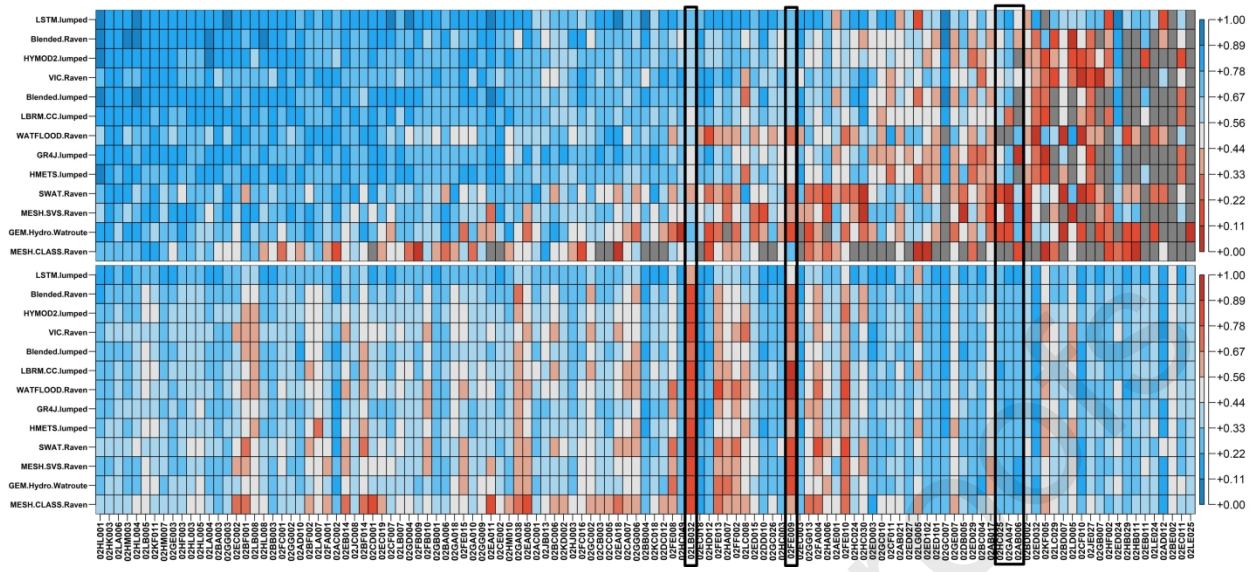




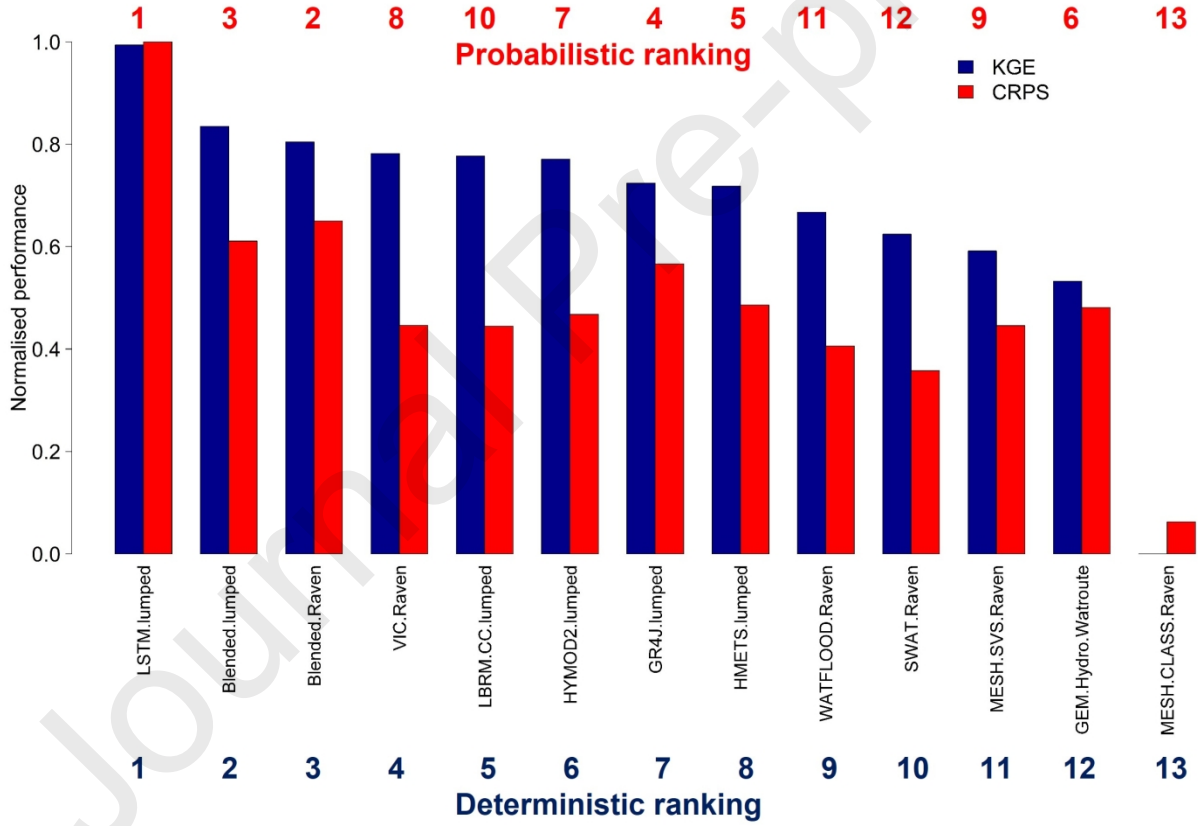
748



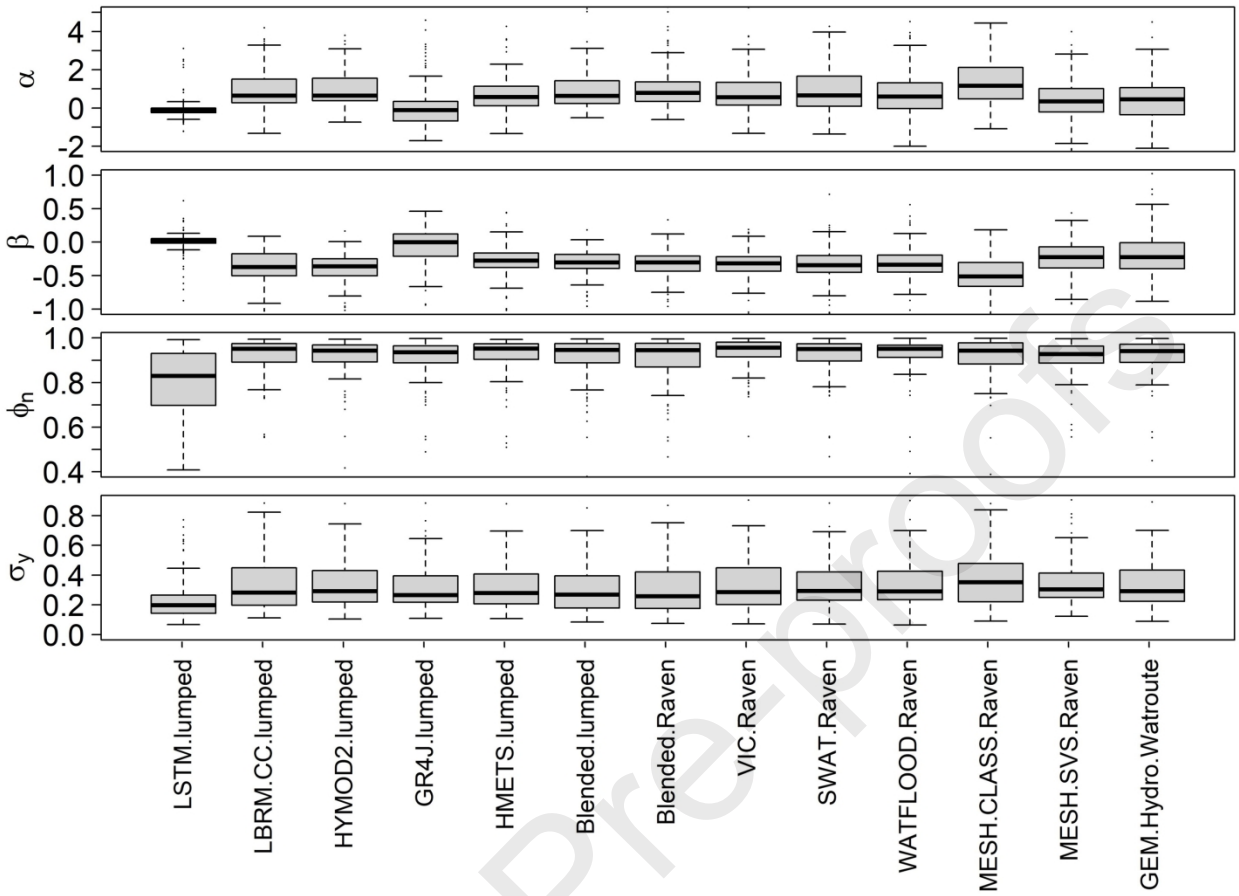
749



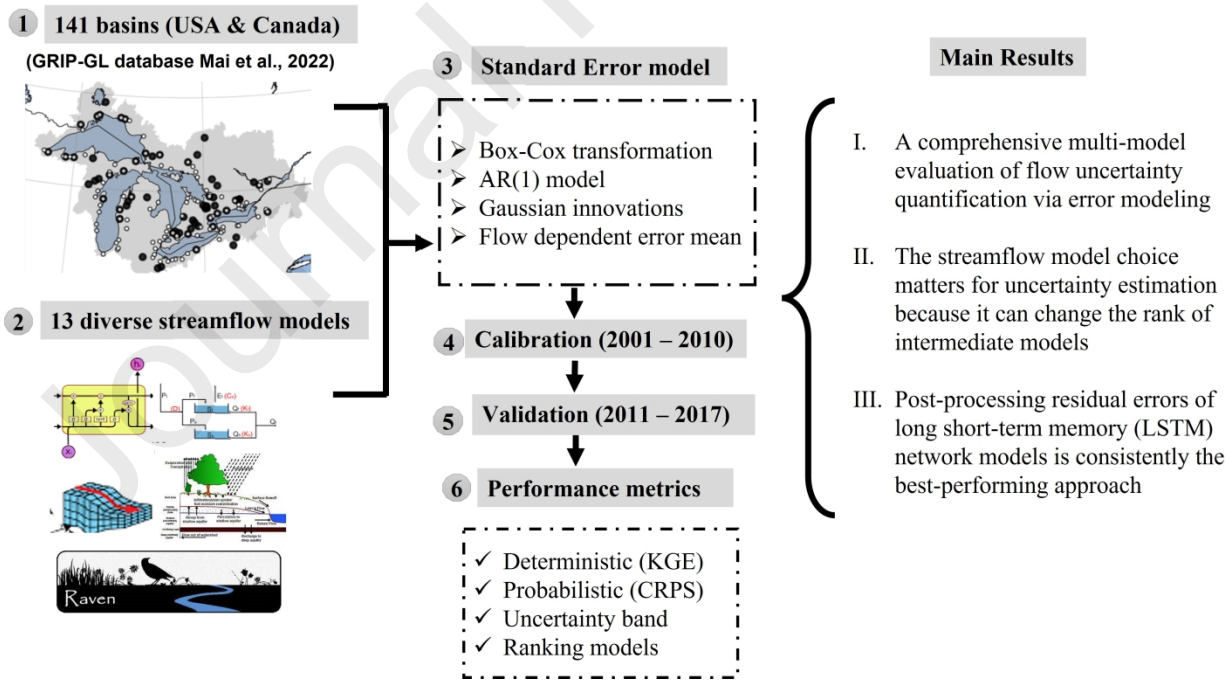
750



751


α
β
ϕ_n
σ_y
LSTM.lumped
LBRM.CC.lumped
HYMOD2.lumped
GR4J.lumped
HMETS.lumped
Blended.lumped
Blended.Raven
VIC.Raven
SWAT.Raven
WATFLOOD.Raven
MESH.CLASS.Raven
MESH.SVS.Raven
GEM.Hydro.Watroute

752


1 141 basins (USA & Canada)
(GRIP-GL database Mai et al., 2022)
2 13 diverse streamflow models
Raven
3 Standard Error model
Box-Cox transformation
AR(1) model
Gaussian innovations
Flow dependent error mean
4 Calibration (2001 – 2010)
5 Validation (2011 – 2017)
6 Performance metrics
Deterministic (KGE)
Probabilistic (CRPS)
Uncertainty band
Ranking models
Main Results
I. A comprehensive multi-model evaluation of flow uncertainty quantification via error modeling
II. The streamflow model choice matters for uncertainty estimation because it can change the rank of intermediate models
III. Post-processing residual errors of long short-term memory (LSTM) network models is consistently the best-performing approach

753

Highlights:

- Uncertainty assessments were estimated in 141 basins for 13 hydrologic model predictions
- Quality of estimates was loosely correlated to model accuracy
- LSTM models were consistently best performing in terms of KGE and CRPS

CRediT authorship contribution statement

Jonathan Romero-Cuellar: Conceptualisation, Investigation, Formal analysis, Visualisation, Software, Methodology, Writing - original draft. **Rezgar Arabzadeh:** Software, Data curation, Visualisation, Formal analysis. **James R. Craig:** Conceptualisation, Methodology, Writing - review & editing, Supervision, Resources. **Bryan A. Tolson:** Conceptualisation, Methodology, Writing - review & editing, Supervision. **Juliane Mai:** Writing - review & editing.

Abstract

Probabilistic streamflow predictions are valuable tools for predictive uncertainty estimation, hydrologic risk management, and support for decision-making in water resources. Usually, predictive uncertainty quantification is developed and assessed using only a single hydrological model, making it difficult to generalize to other model configurations. To tackle this issue, we assess changes in the model performance ranking of diverse streamflow models by applying a residual error model post-processing approach to multiple basins and multiple models. This assessment employed 141 basins from the Great Lakes watershed covering the USA and Canada, and 13 diverse streamflow models, which are evaluated using deterministic and probabilistic performance metrics. As the first study to implement probabilistic methods to diverse streamflow models applied to a multitude of basins, the analysis here examines the dependence of probabilistic streamflow estimation quality on model quality. Our findings show that streamflow model choice influences the robustness of probabilistic predictions. It was found that moving from deterministic to probabilistic predictions using a post-processing approach does not change the streamflow model performance ranking for the best and worst deterministic models, but models of intermediate rank in deterministic evaluation do not have consistent ranking when evaluated in probabilistic mode. Post-processing residual errors of long short-term memory (LSTM) network models are consistently the best-performing model in terms of deterministic and probabilistic metrics. This study highlights the significance of combining deterministic streamflow model predictions with residual error models for improving the quality and increasing the value of hydrological predictions, quantifying uncertainty, and facilitating decision-making in operational water management. It also clarifies the degree to which

probabilistic predictions depend upon good model performance and can compensate for poor model performance.