Hydrol. Earth Syst. Sci., 30, 6907–6925, 2026
https://doi.org/10.5194/hess-30-6907-2026
© Author(s) 2026. This work is distributed under
the Creative Commons Attribution 4.0 License.

Hydrology and Earth System Sciences
Open Access
EGU

Simple Box-Cox probabilistic models for hourly streamflow predictions

Cristina Prieto[1,2,3], **Dmitri Kavetski**[4,1], **Fabrizio Fenicia**[3], **James W. Kirchner**[2,5,6], **David McInerney**[4], **Mark Thyer**[4], **and César Álvarez**[1]

[1]IHCantabria – Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Santander, Spain
[2]Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland
[3]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[4]School of Civil Engineering and Construction Management, Adelaide University, Adelaide, SA, Australia
[5]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[6]Swedish University of Agricultural Sciences (SLU), Umeå, Sweden

Correspondence: Cristina Prieto (prietoc@unican.es)

Received: 27 January 2026 – Discussion started: 4 February 2026
Revised: 21 July 2026 – Accepted: 29 July 2026 – Published: 9 November 2026

Abstract. The increasing availability of hourly scale hydrological data offers valuable benefits for advancing our understanding of catchment processes and improving prediction capabilities. This work contributes to streamflow predictions at the hourly scale by investigating practical methods for uncertainty quantification using probabilistic models. We examine common approaches for representing the heteroscedasticity of streamflow errors using the Box-Cox (BC) transformation and common approaches for representing the autocorrelation of streamflow errors using autoregressive (AR) models. Case studies based on 7 catchments from Spain, Switzerland and USA that cover humid to semi-arid conditions are reported. The results favor Box-Cox transformations with power parameter values of 0–0.5. Notably the Log transformation achieves the best statistical reliability of predictions, while its precision and volumetric bias are not statistically significantly worse than for other transformations considered in this study. Catchment-by-catchment variability becomes more pronounced when performance is stratified to the top 5 % of flows. The results also tend to favor the AR2 and AR3 models over the AR1 model for representing the error autocorrelation, with the addition of moving average terms generally providing little additional benefit. Our findings are broadly consistent with earlier work with daily data, and provide practical guidance for hydrologists and practitioners requiring uncertainty quantification at the hourly scale. We also report progress towards "seamless" aggregation from hourly to longer scales, which is a capability that is desirable in many practical operational contexts.

1 Introduction

The increasing accessibility of hourly hydrological data, in addition to widely available daily time series, offers important benefits for advancing our scientific understanding of catchment processes and improving operational prediction capabilities. High-resolution data enhance our ability to analyze catchment processes and improve predictive models for practical applications. For instance, flash floods, an increasingly serious threat across many regions due to climate change (Yin et al., 2023), demand a detailed understanding of underlying hydrological dynamics and the development of dedicated prediction tools.

Representations of wetting and drying dynamics benefit greatly from hourly data to calibrate and evaluate model behaviour. For example, response to precipitation is highly sensitive to antecedent wetness conditions (Kirchner, 2003; Westerberg and McMillan, 2015; Zheng et al., 2023), which in some catchments can vary very rapidly on sub-daily timescales. On the other hand, hydrological modelling at hourly rather than daily time scales entails additional challenges, as both processes and uncertainties become increasingly complex at finer time resolution. Notably, the dominant

controls on catchment behaviour shift with temporal resolution (Kavetski et al., 2011), so that the key processes affecting short-term predictions are often different to those affecting streamflow at longer time scales. At longer time scales, predictions are shaped more by storage-release processes and general mass balance considerations, whereas hourly predictions require resolving processes such as overland flow, channel routing, and interception dynamics. These dynamics introduce added complexity and typically demand more detailed model structures and data (e.g., see Bieroza et al., 2023; Kirchner et al., 2004). From a more applied perspective, peak flows may last only a few hours, so daily scale predictions can greatly underestimate flood peaks (Ayzel and Heistermann, 2021; Dieter and Arns, 2003). Hourly scales are particularly important in small and mesoscale catchments (with areas less than about $10\,000\,km^2$) due to faster response, e.g., causing "flash floods" (e.g. Forte et al., 2025).

This work contributes to streamflow prediction and uncertainty quantification at the hourly scale. We focus on practical methods for probabilistic prediction, in order to provide a characterization of predictive uncertainty that is accessible to hydrologists and practitioners across a wide range of applications. The pragmatic focus on (relatively) simple approaches is intended to overcome an arguably common reticence among practitioners to use probabilistic modelling methods (e.g., see Hunter et al., 2021).

We consider conceptual hydrological models, which offer the potential to balance accuracy and parsimony in the description of hydrological processes with manageable data requirements and affordable computational cost. Conceptual models such as GR4J, PDM, HBV, VIC and others have been widely used in modelling catchment-scale runoff-generation at daily scales. The GR4H model (Mathevet, 2005), an hourly variant of the GR4J daily model (Perrin et al., 2003) developed empirically over many catchments, has been implemented in many studies worldwide (e.g., de Boer-Euser et al., 2017; Li et al., 2017; van Esse et al., 2013) and showed good efficiency in hourly streamflow predictions. Recent studies on hydrological modelling at hourly scales have also highlighted the promise of machine learning approaches, particularly Long Short Term Memory (LSTM) models (e.g., Gauch et al., 2021). Importantly, the uncertainty estimation methods used in this study are broadly applicable to multiple hydrological modelling approaches, including machine learning, physically based modelling, and conceptual modelling. Here we confine our analysis to conceptual modelling in the interests of simplicity and transparency.

Sources of predictive uncertainty in hydrological modelling include data errors and model approximations (e.g., Clark et al., 2008; McMillan et al., 2011; Prieto et al., 2021; Renard et al., 2011). These uncertainties manifest as differences between model predictions and observed streamflow, which are commonly referred to as residual errors. It is well known that residual errors are typically heteroscedastic, i.e., larger errors in larger flows (e.g., McInerney et al., 2017; Sorooshian and Dracup, 1980), autocorrelated, i.e. multiple consecutive errors with the same sign and similar magnitude (e.g., Evin et al., 2013), and often biased and non-stationary (e.g., Westra et al., 2014). The representation of these characteristics has received significant attention in the hydrological literature, especially at daily time scales (Li et al., 2016; McInerney et al., 2018; McInerney et al., 2017; Wani et al., 2019). Reliable uncertainty quantification provides important practical benefits – in particular, e.g. McInerney et al. (2024); demonstrated that neglecting hydrological model errors can lead to severe underestimation of risk when evaluating water resources system performance.

Compared to daily predictions, hourly predictions are likely to be characterized by stronger heteroscedasticity, bias, autocorrelation, and non-stationarity (Ammann et al., 2019). Accounting for these characteristics in a residual error model is essential for robust model predictions.

There are several approaches for implementing residual error modelling. In the postprocessor approach, the residual error model is analyzed separately given a fixed (precalibrated) hydrological model, with the error model parameters estimated separately from hydrological model parameters (Evin et al., 2014; Li et al., 2016; McInerney et al., 2018; Schoups and Vrugt, 2010). This approach is particularly common because of its flexibility and robust practical performance when used with conceptual models (e.g., Evin et al., 2014; Hunter et al., 2021), and has also been employed with LSTM models (Romero-Cuellar et al., 2024). By contrast, joint inference attempts to estimate error model parameters simultaneously with the hydrological model parameters. Recent applications of joint inference include conceptual models (Ammann et al., 2019) and LSTM models (Klotz et al., 2022). However, while theoretically more appealing, joint inference can suffer from poor identifiability and high computational costs (Evin et al., 2014; Li et al., 2016).

The post-processor approach has been primarily applied at daily time scales. McInerney et al. (2017) recommended transforming streamflow using a Box and Cox (1964) transformation with power parameter of 0.2 or 0.5 to account for error heteroscedasticity (also referred to as "variance stabilisation"), followed by the application of a first order autoregressive model (AR1) to account for error autocorrelation. Follow-up work by Hunter et al. (2021) examined bias correction of previously calibrated hydrological models in simple uncertainty quantification scenarios. The "Error Reduction and Representation in Stages" approach (ERRIS), in addition to heteroscedasticity and bias correction, accounted for differences in autocorrelation between the rising and falling hydrograph limbs (Li et al., 2016). Koutsoyiannis and Montanari (2022) proposed the "Brisk Local Uncertainty Estimator for Generic Simulations and Predictions" method (BLUECAT) which uses empirical distributions of the current predictions to transform a deterministic model into a stochastic predictor with uncertainty assessment.

Hourly predictions and their uncertainty quantification have received relatively less attention, mainly because high-quality hourly data have been scarce until recently, and because prediction uncertainty at this time scale is harder to characterize.

Studies in uncertainty quantification at the hourly scale include joint inference and post-processor approaches. Ammann et al. (2019) employed joint inference where residual errors were characterised accounting for heteroscedasticity, right skew due to non-negativity of streamflow, excess kurtosis (fat tails), and reduced autocorrelation during wet periods. The Ammann et al. (2019) error model offers the potential to provide a comprehensive description of predictive uncertainty, but its practical limitations include a large number of parameters, which are difficult to estimate particularly in a joint inference setup. Li et al. (2017) adapted the ERRIS approach to hourly predictions by allowing different mixtures of Gaussian distributions for the rising and falling hydrograph limbs. In Li et al. (2021), the ERRIS approach was extended further by treating zero streamflow as censored data, which is beneficial in ephemeral catchments (see also McInerney et al., 2019).

The treatment of autocorrelation in hourly residuals has also received attention. The studies by Li et al. (2016), Li et al. (2017), Li et al. (2021) and Ammann et al. (2019) limited their attention to an AR1 model, though Ammann et al. (2019) reported that higher orders for the autoregressive models might be needed. Wani et al. (2019) proposed using copulas to separate the specification of the dependence structure and the marginal distribution of the residuals. For example, negative streamflow predictions can be avoided by selecting a marginal distribution with corresponding support. The dependence structure can be controlled by the choice of copula, for example allowing for stronger dependence of errors during low vs high flows. Limitations include poor identifiability when copula parameters interact with hydrological model parameters and heteroscedasticity parameters Wani et al. (2019).

Our review of the current literature suggests unexplored opportunities in the design of simple and practical approaches for incorporating uncertainty in hourly streamflow predictions, particularly when using conceptual hydrological models.

The comprehensive error models used in many existing hourly formulations (e.g., Ammann et al., 2019; Wani et al., 2019) are theoretically appealing, but in practice lead to complex likelihood functions and poor identifiability/stability, especially in joint inference setups (e.g., see Ammann et al., 2019).

Postprocessing methods have shown practical success at daily time scales, both with conceptual models and LSTM models (Hunter et al., 2021; Klotz et al., 2022; Li et al., 2021; McInerney et al., 2017; Romero-Cuellar et al., 2024), but have not yet been sufficiently tested at hourly time scales. Simple postprocessor approaches, especially those using the well-known Box-Cox transformation, remain largely unexplored. Some studies, e.g., Klotz et al. (2022), account for heteroscedasticity but not for autocorrelation.

So far, studies at hourly time scales considered only a single streamflow transformation for heteroscedasticity and only a single model for error autocorrelation, and have not reported performance when aggregating flows from hourly to daily and monthly time scales, nor for special conditions such as high flows.

Prediction performance can vary substantially depending on flow magnitude and regime. While flow stratifications have been considered in previous studies, they have generally been limited to deterministic model analyses (Addor et al., 2017; Blöschl et al., 2019; Kirchner, 2003; Nevo et al., 2022; Paltan et al., 2017; Prieto et al., 2020; Prieto et al., 2024), with some exceptions being McInerney et al. (2021) where flow stratified performance is reported for probabilistic forecasts at the daily scale.

Another practical aspect of hydrological prediction that has received recent attention is the ability to achieve "seamless" prediction, including predictions that remain reliable when aggregated to coarser time scales, e.g., from daily to monthly as demonstrated in McInerney et al. (2020). Such predictions, if available, avoid the need for multiple models and prediction products, and thus are beneficial in many practical applications (e.g., McInerney et al., 2022).

This study evaluates simple approaches for representing uncertainty in hourly streamflow predictions. Our specific aims are:

1. Quantify the uncertainty in conceptual hydrological models for hourly streamflow predictions.

2. Recommend residual error models for practical applications by comparing several heteroscedastic and autoregressive residual error models, using multiple statistical performance metrics (reliability, precision and bias).

3. Explore additional aspects of the performance of the error model, namely

    i. Performance for the top 5 % of flows (stratified flow performance).

    ii. Performance at time scales aggregated from hourly to daily and monthly.

For error heteroscedasticity we consider the Box-Cox transformation with several common values of the power parameter, using methods similar to the earlier daily scale work by McInerney et al. (2017). For error autocorrelation we consider several Autoregressive (AR) and Autoregressive Moving Average (ARMA) models.

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 case study includes 7 catchments from Spain, Switzerland and USA that cover humid to semi-arid conditions.

2 Theoretical development

2.1 Basic definitions

Let $q_t^{\boldsymbol{\theta}_h}$ denote a streamflow prediction at time step t obtained using a deterministic hydrological model h with parameters $\boldsymbol{\theta}_h$ and inputs $\mathbf{x}$ (up to step t),

$$q_t^{\boldsymbol{\theta}_h} = h(\boldsymbol{\theta}_h; \mathbf{x}_{1:t}) \tag{1}$$

A probabilistic model of streamflow at time t, $Q_t(\boldsymbol{\theta}; \mathbf{x}_{1:t})$, with probability density function (pdf) $p(q_t|\boldsymbol{\theta}, \mathbf{x}_{1:t})$ and general parameters $\boldsymbol{\theta}$ is formulated next in order to represent the predictive uncertainty due to residual errors, which are intended to represent the combined effect of data and model errors. This notation distinguishes the random variable Q_t from a realization q_t. To reduce clutter, we will drop the conditioning on $\mathbf{x}$ as it is common to all cases.

2.2 Residual error model

In this work, residual error models are constructed by combining a streamflow transformation to represent error heteroscedasticity and an autoregressive model to represent error autocorrelation. These components are described next.

2.2.1 Box-Cox transformation to represent error heteroscedasticity

Consider the streamflow distribution

$$z(Q_t; \boldsymbol{\theta}_z) \sim \mathcal{N}\left(z(q_t^{\boldsymbol{\theta}_h}; \boldsymbol{\theta}_z); \sigma_\eta^2\right) \tag{2}$$

This probabilistic model corresponds to an additive Gaussian residual error model in transformed space,

$$z(Q_t; \boldsymbol{\theta}_z) = z(q_t^{\boldsymbol{\theta}_h}; \boldsymbol{\theta}_z) + \eta_t \tag{3}$$

$$\eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{4}$$

where $z(q)$ is a transformation function with parameters $\boldsymbol{\theta}_z$ and η_t is the normalized residual error. It is assumed that η_t follows a zero-mean Gaussian distribution with variance σ_η^2.

The full parameter set of the probabilistic model is $\boldsymbol{\theta} = \left\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_z, \boldsymbol{\theta}_\eta\right\}$ and includes additional parameters describing the transformation function z and properties of normalized residuals $\boldsymbol{\eta}$.

We use the Box-Cox transformation (Box and Cox, 1964) with parameters $\boldsymbol{\theta}_z = \{\lambda, A\}$, where λ is the power parameter and A is an offset parameter.

$$z(q; \boldsymbol{\theta}_z) = z(q; \lambda, A) = \begin{cases} \frac{(q+A)^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \log(q + A) & \text{if} \quad \lambda = 0 \end{cases} \tag{5}$$

The value $\lambda = 0.2$ is recommended based on work with daily streamflow predictions (McInerney et al., 2017); in this paper we revisit this question using hourly data. The offset parameter A is typically set to a small value to avoid numerical problems when $q \approx 0$.

2.2.2 ARMA models to represent error autocorrelation

The normalized residuals are assumed to follow an autoregressive (AR) model of order N_ϕ,

$$\eta_t = \sum_{i=1}^{N_\phi} \phi_i \eta_{t-i} + y_t \tag{6}$$

where y_t is the innovation (random component or "noise" term) at time t and $\boldsymbol{\phi} = \left\{\phi_i; i = 1, \ldots, N_\phi\right\}$ are the autoregressive coefficients.

The innovations are assumed to follow a Gaussian distribution with a mean of zero and standard deviation σ_y^2,

$$y_t \sim \mathcal{N}(0, \sigma_y^2) \tag{7}$$

The parameters of the residual error model are then $\boldsymbol{\theta}_\eta = \{\sigma_y, \boldsymbol{\phi}\}$.

We also consider a more general, autoregressive moving average ARMA model of order (N_ϕ, N_φ) dependent on N_ϕ past residuals and N_φ past innovations,

$$\eta_t = \sum_{i=1}^{N_\phi} \phi_i \eta_{t-i} + \sum_{j=1}^{N_\varphi} \varphi_j y_{t-j} + y_t \tag{8}$$

where $\boldsymbol{\varphi} = \{\varphi_j; j = 1, \ldots, N_\varphi\}$ are the moving average (MA) parameters so that $\boldsymbol{\theta}_\eta = \left\{\sigma_y, \boldsymbol{\phi}, \boldsymbol{\varphi}\right\}$.

2.3 Postprocessor approach for parameter estimation

The parameters $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\theta}}_h, \hat{\boldsymbol{\theta}}_\eta\}$ are estimated from observed streamflow time series $\tilde{\mathbf{q}} = \{\tilde{\mathbf{q}}_t; t = 1, \ldots, N_t\}$ where N_t is the total number of time steps. We employ a two-stage postprocessor approach similar to McInerney et al. (2018): we used the Least Squares method for Stage 1 and the Maximum Likelihood method for Stage 2, as described next.

2.3.1 Stage 1: Calibration of the deterministic hydro model parameters

Stage 1 calibrates the hydrological model parameters $\boldsymbol{\theta}_h$ using the Least Squares (LS) objective function, which is applied in

transformed space with fixed transformation parameters $\boldsymbol{\theta}_z$,

$$\Phi_{\text{obj}}(\boldsymbol{\theta}_h) = \Phi_{\text{SSE}}\left(z(\mathbf{q}[\boldsymbol{\theta}_h]; \boldsymbol{\theta}_z); z(\tilde{\mathbf{q}}; \boldsymbol{\theta}_z)\right) = \sum_{t=1}^{N_t}\left(z(\tilde{q}_t; \boldsymbol{\theta}_z) - z(q_t; \boldsymbol{\theta}_z)\right)^2 \quad (9)$$

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h}{\text{argmin}}\,\Phi_{\text{obj}} \quad (10)$$

As shown in McInerney et al. (2018), this approach represents a pragmatic implementation of Maximum Likelihood estimation under the assumption of Gaussian errors.

2.3.2 Stage 2: Residual error model calibration

Stage 2 estimates the parameters of the residual error model $\boldsymbol{\theta}_\eta$. The residual error is calculated as

$$\tilde{\eta}_t = z(\tilde{q}_t; \boldsymbol{\theta}_z) - z(\hat{\boldsymbol{\theta}}_h; \boldsymbol{\theta}_z) \quad (11)$$

using the hydrological model parameters estimated in Stage 1.

The AR and ARMA model parameters are estimated using the Maximum Likelihood method (Box et al., 1994; Hamilton, 1994) implemented in the econometrics toolbox in Matlab.

Note that we specify a residual error model with zero mean, see Eq. (7). This model assumption was verified empirically by confirming that the mean of the estimated innovations (and residuals) is close to zero. Residual error models with non-zero mean have also been suggested in the literature (e.g., Li et al., 2017), and may be required if different transformations are applied in Stages 1 and 2 (see Hunter et al., 2021).

2.4 Generation of predictive distributions

The predictive distribution of streamflow is given by the probabilistic model $Q_t(\hat{\boldsymbol{\theta}}; \mathbf{x}_{1:t})$, i.e. using the combined hydrological model and residual error model with estimated parameters $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\theta}}_h, \hat{\boldsymbol{\theta}}_\eta\}$ and inputs $\mathbf{x}$.

Given an already computed deterministic streamflow prediction $q_t^{\hat{\boldsymbol{\theta}}_h}$, a predictive replicate $q_t^{\text{pred}(r)}$ for the case of an error model with AR(N_ϕ) autocorrelation structure is generated as follows.

1. Sample innovations from a Gaussian distribution

$$y_t^{(r)} \leftarrow \mathcal{N}(0, \hat{\sigma}_y^2) \quad (12)$$

2. Calculate the residuals using

$$\eta_t^{(r)} \leftarrow \sum_{i=1}^{N_\phi} \hat{\phi}_i \eta_{t-i}^{(r)} + y_t^{(r)} \quad (13)$$

3. Apply the inverse transformation

$$q_t^{\text{pred}(r)} = z^{-1}\left(z(q_t^{\hat{\boldsymbol{\theta}}_h}; \boldsymbol{\theta}_z) + \eta_t^{(r)}\right) \quad (14)$$

Here, we impose a lower bound $q_{\min}$ of zero and an upper bound $q_{\max}$ of 10 times the maximum observed streamflow (Wang et al., 2012).

4. The complete set of all replicates, i.e. the predictive distribution, is

$$\mathbf{q}^{\text{pred}} = \left\{q_t^{\text{pred}(r)}; t = 1, \ldots, N_t; r = 1, \ldots, N_r\right\} \quad (15)$$

where N_r is the number of replicates.

3 Case study

This section describes the case study catchments and methods. Note that the focus of this study is on uncertainty quantification via probabilistic predictions. For this reason, all model performance evaluations undertaken in this section refer to the statistical performance of the probabilistic model, which as defined in Eq. (2) includes the deterministic hydrological model and the residual error model. Moreover, as the deterministic hydrological model is kept fixed, our comparison focuses on differences in the performance of the residual error models. The subsections below provide full details.

3.1 Catchments and observational data

A total of 7 catchments are used in this study, namely Lasarte (Spain), Wängi (Switzerland), Smith River (CA, USA), Gila (NM, USA), French Broad (NC, USA), Baron Fork (OK, USA), San Francisco (AZ, USA). These catchments span humid to arid conditions, according to the Arora (2002) classification based on the PET / P ratio; see Table 1 for details. Hourly time series of rainfall, streamflow and potential evapotranspiration are employed. To avoid additional modelling complications, we exclude catchments with ephemeral flow and/or snow-dominated hydrology.

3.2 Deterministic hydrological model

The conceptual hydrological model GR4H (Génie Rural à 4 paramétres Horaires) is employed for the deterministic component in Eq. (1). GR4H is a member of the GR series of models, which have been applied in a wide range of catchments including in Australia, France, and Switzerland (e.g., Coron et al., 2012; Dal Molin et al., 2020; Perrin et al., 2003; van Esse et al., 2013).

GR4H is the hourly variant of the daily GR4J rainfall-runoff model (Perrin et al., 2003), retaining the same overall structure but using different values for several fixed (internal) parameters. The model comprises two conceptual reservoirs: the production store, which governs evapotranspiration and determines the effective portion of rainfall that contributes to runoff, and the routing store, which regulates baseflow generation. Additionally, two lag functions control the timing and shape of the hydrograph peak. A schematic of the GR4 model structure is given in Fig. 1 of Perrin et al. (2003).

Table 1. Summary of case study catchments and their properties.

Catchment	Location	Area, km^2	Mean altitude, m	Mean annual flow, $mm\,yr^{-1}$	PET / P	Q / P	Classification (Arora 2002)	Data source
Lasarte	Basque Country Spain	861	837	989	0.51	0.52	Humid	Basque Water Agency (URA)
Wängi	Wängi Switzerland	80.15	592.14	728	0.50	0.60	Humid	Swiss Federal Office for the Environment (FOEN)
Smith River	California, USA	614	24.2	2125	0.28	0.73	Humid	Model Parameter Estimation Experiment (MOPEX) (Duan et al., 2006)
Gila	New Mexico, USA	1864	1418.8	28	2.22	0.09	Semi-arid	MOPEX
French Broad	North Carolina, USA	945	594.4	708	0.43	0.59	Humid	MOPEX
Baron Fork	Oklahoma, USA	312	213.7	490	0.99	0.3	Sub-humid	MOPEX
San Francisco	Arizona, USA	2763	1047.3	20	2.33	0.08	Semi-arid	MOPEX

GR4H has four calibration parameters: the capacity of the production store θ_1 (mm), the groundwater exchange coefficient θ_2 ($mm\,h^{-1}$) that accounts for groundwater import or export, the capacity of the routing store θ_3 (mm), and the time base of the unit hydrograph θ_4 (h).

3.3 Residual error model

3.3.1 Streamflow transformations to represent error heteroscedasticity

We consider four fixed values of the Box-Cox power parameter:

- $\lambda = 0$, equivalent to the logarithmic transformation, here denoted "Log",
- $\lambda = 0.2$, recommended at daily scale in McInerney et al. (2017); denoted "BC02",
- $\lambda = 0.5$, square root transformation, denoted "BC05", and
- $\lambda = 1$, i.e., no transformation, corresponding to the Simple Least Squares method, denoted "SLS".

The offset parameter A is fixed at 10 % of the mean flow in the catchment being modelled (McInerney et al., 2018).

3.3.2 Autoregressive models to represent error autocorrelation

We consider three autoregressive error models: AR1, AR2 and AR3. In addition, we consider the inclusion of moving average terms: MA1, MA2 and MA3.

3.4 Split sample validation

A split sample validation approach is employed. In a given catchment, the data is split into two time periods. The (probabilistic) model is calibrated on the first period and used to generate (probabilistic) streamflow predictions for the second period. The model is then calibrated on the second period and used to generate streamflow for the first period. The streamflow predictions from the two "validation" periods are then concatenated into a single long time series of the length of the full data set. A model warmup period is employed, comprising 2–3 years preceding the calibration period.

3.5 Flow stratification and aggregation

The performance of the probabilistic predictions in each catchment is evaluated for the following data sets:

1. All flows, i.e., the entire data period as (by definition) it includes the full range of streamflow magnitudes.

2. Top 5 % of flows. This performance evaluation is included because many hydrological applications focus on higher flows (Addor et al., 2017). Alternative definitions of "high" flows, such as the top 2 %–10 % of flows, have also been used in the literature (United States Environmental Protection Agency, 2007; Yilmaz et al., 2008). In this study, the streamflow threshold defining the top 5 % of flows in a validation period is set according to the (deterministic) model streamflow in the associated calibration period. More specifically, with reference to the split sample validation setup (Sect. 3.4), the top 5 % of flows in the (concatenated) validation period is defined as the combined subsets of model streamflow in periods 1 and 2 that exceed the 95th percentile of model streamflow in periods 2 and 1 respectively. This stratification approach accounts for potentially different streamflow characteristics in the calibration and validation periods.

3. Aggregated streamflow time series. This performance evaluation is included given the recent interest in "seamless" streamflow predictions, where a single model at a fine resolution is used to generate predictions at multiple coarser time scales (McInerney et al., 2020). Here, we consider aggregations from hourly to daily and monthly scales. For example, daily streamflow is obtained by aggregating hourly streamflow for each hour of the (calendar) day. This aggregation approach is distinct from the common approach of constructing a streamflow model directly at the daily scale. Similarly, monthly streamflow is obtained by aggregating together hourly streamflow for each hour of the calendar month.

3.6 Performance comparison and evaluation approach

The probabilistic predictions and their key components are evaluated as follows:

1. *Comparison of transformations employed within the residual error model at the hourly scale*. This step compares the 4 transformation setups listed in Sect. 3.3.1, namely Log, BC02, BC05 and SLS. The statistical reliability, precision and bias of the probabilistic streamflow predictions are evaluated using the performance metrics summarised in Sect. 3.6.1, with full definitions given in Appendix A. Note that these metrics are commonly used in studies on probabilistic prediction in hydrology (see, e.g., Evin et al., 2014; Hunter et al., 2021; Renard et al., 2010; and many others), the implications of this choice of evaluation metrics are discussed later in Sect. 5.1.

   The statistical significance of differences in the performance metrics for different transformations is then tested at the 95 % confidence level using the Wilcoxon test, following a similar approach to McInerney et al. (2017).

   This comparison is carried out for the two flow stratifications listed in Sect. 3.5, namely for all flows and for the top 5 % of flows.

   Note that the reliability, precision and volumetric bias metrics are applied to the marginal predictive distribution of streamflow, which in this study refers to the unconditional distribution of streamflow predictions at a series of time steps, without an attempt to condition on observations at preceding time steps. These marginal predictive distributions are not affected by the autocorrelation structure of the residuals, because (i) all ARMA models generate time series of random variables (here, residual errors) with the same marginal distribution at each time step; and (ii) the two-stage postprocessor approach in Sect. 2.3 disregards posterior parameter uncertainty. Under this setup, the comparison of the transformations in this part of the analysis can proceed for a single autoregressive model, which substantially simplifies the analysis. The results reported here are computed using the AR3 model, and it was confirmed empirically that the use of AR1 and AR2 does not change the reliability, precision and bias metrics (as expected from the theoretical considerations above).

2. *Comparison of autoregressive models within the residual error model at the hourly scale*. This step compares the AR and ARMA models listed in Sect. 3.3.2 fitted to the residuals obtained from the hydrological model with parameter values estimated in Stage 1, for all transformations. The comparison is carried out using the partial autocorrelation of the innovations, as described in Sect. 3.6.3 below, seeking to identify the autoregressive model that yields the least autocorrelated innovations (i.e., removes the most autocorrelation from the residuals).

3. *Comparison of probabilistic predictions across multiple time scales from daily to aggregated*. This step considers all residual error models (all combinations of transformations and autoregressive models) in the context of seamless predictions, where a single streamflow prediction product can be aggregated to the scale of interest in a particular application. In this analysis, when comparing individual residual error models, performance metrics are averaged over all catchments. When comparing residual error performance across time scales, the performance metrics are additionally averaged over all autoregressive models, whereas when comparing residual error performance with respect to the autoregressive model, the performance metrics are additionally averaged over all time scales. Details of this averaging procedure are provided in Appendix B. Notably, all performance metrics are dimensionless, which enables averaging without introducing dependence on scale-specific flow ranges, etc.

The three-step evaluation strategy above is selected because it strikes a balance between detailed evaluation of individual components of the probabilistic predictions as well as evaluation of its overall performance across a range of time scales.

Note also that, given that (i) the probabilistic model in this work comprises the deterministic model and the residual error model, and (ii) the same deterministic model (GR4H) is used in all experiments, it follows that the comparisons of predictive performance in this study are in fact comparisons of the residual error models under consideration.

3.6.1 Performance metrics: Reliability, precision and volumetric bias

A probabilistic prediction is considered (statistically) reliable if the observations over a series of time steps are consistent with being samples from the predictive distribution. We quantify reliability using the reliability metric from Renard et al. (2010) and McInerney et al. (2017). This metric is derived from predictive quantile-quantile (PQQ) plots (Laio and Tamea, 2007; Renard et al., 2011; Thyer et al., 2009) and has been used in numerous subsequent studies (e.g., Dal Molin et al., 2023; Klotz et al., 2022; Koutsoyiannis and Montanari, 2022; McInerney et al., 2017; Montanari and Koutsoyiannis, 2025; Vrugt, 2024). Better reliability corresponds to lower values of this metric, with zero representing "perfect" reliability. Precision (often referred to as "sharpness" or "resolution" in the forecasting literature) refers to the width or spread of a probabilistic prediction. Here we use the precision metric from McInerney et al. (2017), see also Hunter et al. (2021), defined as the standard deviation of the predictive distribution averaged over all time steps, scaled by the mean observed flow. Better precision corresponds to lower values of this metric.

Volumetric bias is included to examine the long-term water balance behaviour of the predictions. The volumetric bias metric is taken from McInerney et al. (2017), and is defined as the absolute error in the total predicted volume, averaged across all predicted replicas and accumulated over all time steps, relative to the total volume of the observed streamflow time series. Better volumetric bias corresponds to lower values of this metric.

The full definitions of these performance metrics are given in Appendix A.

3.6.2 Performance of residual error models across multiple metrics and catchments

A single residual error model might not achieve the best performance across all metrics in a given catchment. In addition, the performance of a residual error model may vary across catchments. Hence, we employ a comparison procedure based on the approach of McInerney et al. (2017) to identify the "best-metric" residual error models for each performance metric, as follows:

1. *Identify the "best-median" residual error model*, defined as the residual error model with the best median values of the performance metric. For example, suppose residual error models A, B, C and D have performance metric values m_A, m_B, m_C and m_D, where the notation $m_A = \{m_{A,1}, m_{A,2}, m_{A,3}\}, \ldots$ refers to metric values in catchment 1, 2, 3, etc. Further, suppose median $m_A <$ median $m_B <$ median $m_C <$ median m_D. Then residual error model A is the "best-median" residual error model.

2. *Identify the set of best-metric residual error models*, defined as the residual error models with performance metric values that are statistically similar to those of the best- median residual error model. This is established using the paired Wilcoxon signed-rank test (Bauer, 1972) to check for statistically significant differences at the 95 % confidence level between residual error models A vs. B, A vs. C, A vs. D. The use of a paired test ensures that the metric values are compared case-by-case, in this work catchment by catchment, i.e., residual error model A for catchment 1 is only compared to residual error model B for catchment 1, C for catchment 1 and D for catchment 1.

This procedure is carried out separately for the three performance metrics (reliability, precision and volumetric bias).

3.6.3 Evaluation of ARMA models to represent error autocorrelation

The following procedure is employed to examine autoregressive models in a given catchment:

1. Select an initial set of AR models, here AR1, AR2 and AR3 (Sect. 3.3.2);

2. Apply a residual error model with a selected streamflow transformation;

3. Estimate the innovations by inverting Eq. (6);

4. Estimate the partial autocorrelation function (PACF) of the innovations. Here, we employ the Matlab function "parcorr" (Box et al., 2015; Hamilton, 1994);

5. Compare the AR models to find the model that yields innovations with PACFs closest to zero across all lags. This AR model is taken as the best in representing the autocorrelation of the residual errors;

6. Consider the inclusion of MA components in addition to AR components, and check whether these additional components provide major further improvements in the PACF of the innovations. Here we consider MA1, MA2 and MA3 components (Sect. 3.3.2).

Albeit subjective, this comparison procedure enables us to identify a parsimonious model for capturing the error autocorrelation.

4 Results

4.1 Comparison of residual error models for all flows

Table 2 lists the best-metric residual error models and the best-median residual error models for the individual performance metrics. Note that these residual error models are identified by applying the performance metrics to all flows (i.e., without any stratification). These results are described below.

4.1.1 Reliability

Figure 1a compares the reliability metrics of residual error models. The rankings are generally consistent across the catchments. The Log transformation provides the best-median and best-metric residual error model. The next best transformation is BC02, followed by BC05. Using no transformation ($\lambda = 1$, SLS) yields the worst reliability. These results are similar to the earlier work of McInerney et al. (2017) with daily data, which also favoured the Log, BC02 and BC05 transformations.

The performance metrics of the residual error models in the Lasarte catchment and illustrative time series of probabilistic streamflow predictions are given in Fig. 2. The first column in Fig. 2 shows the PQQ plots for all transformations. The shape of the PQQ plots for BC05 and SLS indicates an overestimated uncertainty when all flows are considered. This behaviour occurs because, even in a perennial catchment the majority of the hours have low flows, i.e., the PQQ plot is dominated by time steps with low flows, for which BC05 and SLS can overestimate uncertainty and produce wider predictive limits than the other transformations. This behaviour is similar to the findings in McInerney et al. (2017).

Figure 2, specifically the hydrographs, also shows that SLS tends on average to under-estimate uncertainty in the high flows, whereas Log and BC02 tend to provide better and more balanced coverage of the observed data.

4.1.2 Precision

Figure 1b compares the precision metrics achieved by the residual error models. The rankings are generally consistent across the catchments. Table 2 and Fig. 1b show that BC02 is the best-median model and BC02 and Log are the best-metric models. BC05 and SLS perform statistically significantly worse than Log and BC02 (see Table 2). Similar to the case for reliability, SLS has the worst precision (see Fig. 1b).

The hydrographs in Fig. 2 show that SLS (i.e. $\lambda = 1$) has narrower predictive limits for high flows (i.e. lower uncertainty) and wider predictive limits for low flows (i.e. higher uncertainty) compared with Log, BC02 and BC05. Further, the bar plots show that precision is worst for SLS.

4.1.3 Volumetric bias

Figure 1c compares the volumetric bias metrics of residual error models. Here the rankings vary appreciably across catchments, notably SLS is the best-performing model in five catchments but the worst-performing model in the Baron and San Francisco catchments. Overall, Fig. 1c and Table 2 show that SLS is the best-median model and the best-metric models are Log, BC02 and BC05. However, Log, BC02 and BC05 are not statistically significantly worse than SLS.

The bar plots in Fig. 2 show that SLS has the least volumetric bias, but as shown in Table 2 the other transformations are not statistically significantly worse when evaluated over the case study catchments.

4.2 Comparison of residual error models for top 5 % of flows

Figure 3 compares the performance metrics of the residual error models for the top 5 % of flows. In contrast to the performance for all flows, the performance for the top 5 % of flows show substantial differences in rankings across the catchments. In terms of the median performance, BC05 has slightly better reliability than the other transformations (Fig. 3a), whereas SLS has slightly better precision (Fig. 3b) and bias (Fig. 3c). However, the differences in reliability and precision are found to be not statistically significant. For volumetric bias, SLS and BC05 are the best-metric transformations: while SLS performs slightly better than BC05, the difference is not statistically significant. Bias also decreases from Log to SLS, with Log and BC02 performing statistically significantly worse than SLS. Note that the three metrics show considerable variability across the catchments. Table 3 summarises the best-metric residual error models and the best-median residual error models for the top 5 % of flows.

4.3 Comparison of autoregressive models to represent the error autocorrelation

Figures 4 and 5 illustrate the performance of the AR1, AR2 and AR3 models in the Smith River and Lasarte catchments respectively. Autoregressive models for which the (estimated) innovations have the PACF closest to zero are considered to provide a better representation of the autocorrelation of residual errors (see Sect. 3.6.3). Overall, autocorrelation is better captured by the AR2 model than by the AR1 model. In some catchments, such as Smith River, the AR3 model captures the autocorrelation better than the AR2 model. For lags above 3, the PACF of the innovations is generally stable around the 95 % confidence bands.

Table 2. Summary of best-median residual error model and best-metric residual error model across all catchments for all flows.

Metric	Best-metric residual error model (best-median residual error model in bold)
Reliability	**Log**
Precision	**BC02**, Log
Volumetric Bias	Log, BC02, BC05, **SLS**

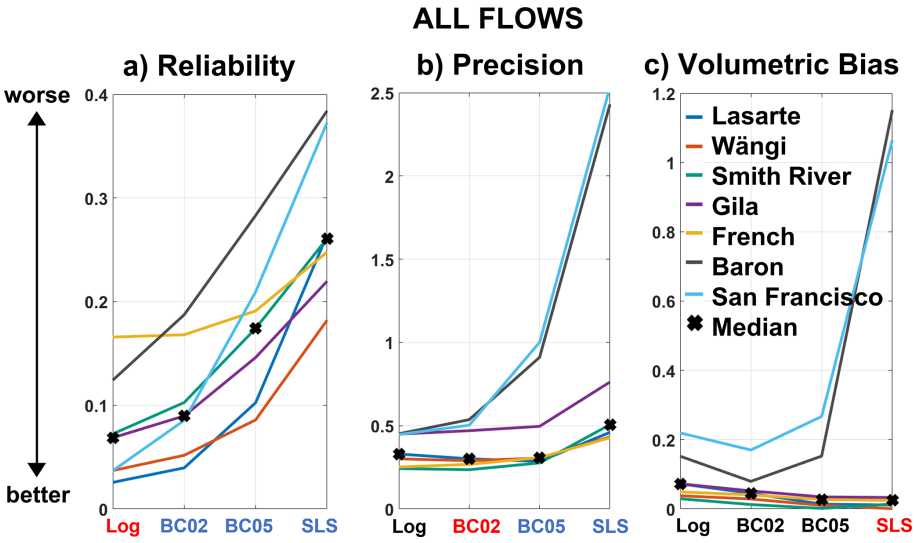


Figure 1. Performance of residual error transformations across all catchments for all flows (i.e., no stratification). Log transformation provides the best reliability. The best-median transformation is indicated in red. The blue x-label indicate transformations that perform statistically significantly worse (95 % confidence) than the best-median transformation in terms of paired Wilcoxon test of the metric values. Bold black means there is no statistically significant difference between the transformation and the best-median transformation.

Figure 6 illustrates the performance of the AR3 model vs ARMA(3,1), ARMA(3,2) and ARMA(3,3) models in the Lasarte catchment. The PACF of the corresponding innovations is similar, indicating that additional MA terms do not provide a large improvement, i.e., the error autocorrelation appears to be effectively captured by the AR3 model.

4.4 Performance at aggregated time scales

The probabilistic predictions are now evaluated under temporal aggregation from hourly to daily and monthly scales. Table 4 presents the following average performance metrics: (i) catchment-average performance metrics for all combinations of transformations and autoregressive models (interior cells of the table); (ii) catchment-average performance metrics at each time scale additionally averaged over the autoregressive models (row labelled "Ave AR"), and (iii) catchment-average performance metrics additionally averaged over the time scales (column labelled "Ave (h, d, m)"). This information is presented for all 4 transformations, in Table 4. Appendix B provides details of the metric averaging procedure.

Table 4 shows that using the Log transformation, model performance when aggregating from hourly to daily is relatively stable and preserves reliability and precision. Further aggregation to monthly scale results in degraded performance in most catchments, typically manifested as a loss of reliability. This trend can also be seen for the average performance across the AR models (bold rows in Table 4).

For example, for the Log transformation, the average reliability and precision barely change from hourly to daily: reliability is 0.079 at hourly time scale and 0.081 at daily; and precision is 0.352 and 0.332, respectively. However, further aggregation to monthly time scale degrades reliability more substantially, to approximately 0.2.

These trends largely hold regardless of the value of λ, though as noted earlier Log and BC02 achieve better absolute performance. The average reliability across the autoregressive models, for each value of λ, shows that both reliability and precision deteriorate from Log to SLS. For example, for AR3, for the Log transformation, reliability is 0.122 and precision is 0.281; while for the SLS, reliability is 0.298 and precision is 0.717.

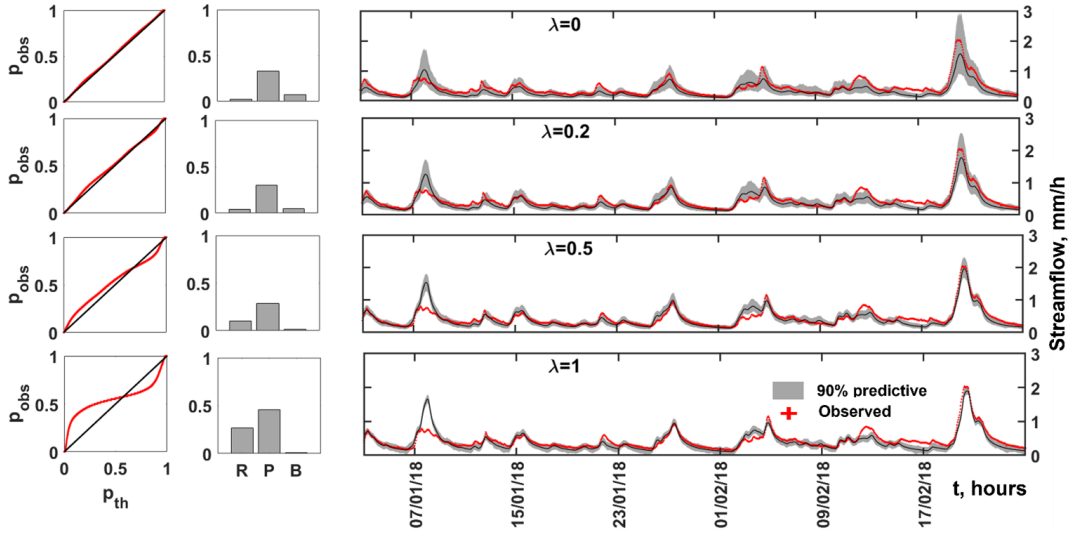


Figure 2. Illustration of probabilistic predictions in the Lasarte catchment. The results for residual error models are arranged row-by-row, for λ from 0 to 1. The left column shows the PQQ plots (Sect. 3.6.1). The bar plots in the middle column show the values of Reliability, Precision, and Volumetric Bias (R, P, B respectively), with lower values corresponding to better performance. The right column shows the hourly hydrographs for a selected representative time period, with red crosses representing observed streamflow, the solid black line indicating the deterministic model prediction, and the grey shading indicating the 90 % prediction limits. Note that the PQQ plots and performance metrics are reported for the entire evaluation time period, not only the shorter time period shown here for illustration purposes. It can be seen that $\lambda = 1$ (SLS) tends on average to under-estimate uncertainty of high flows, whereas $\lambda = 0$–0.2 tend to provide better and more balanced coverage of the observed data.

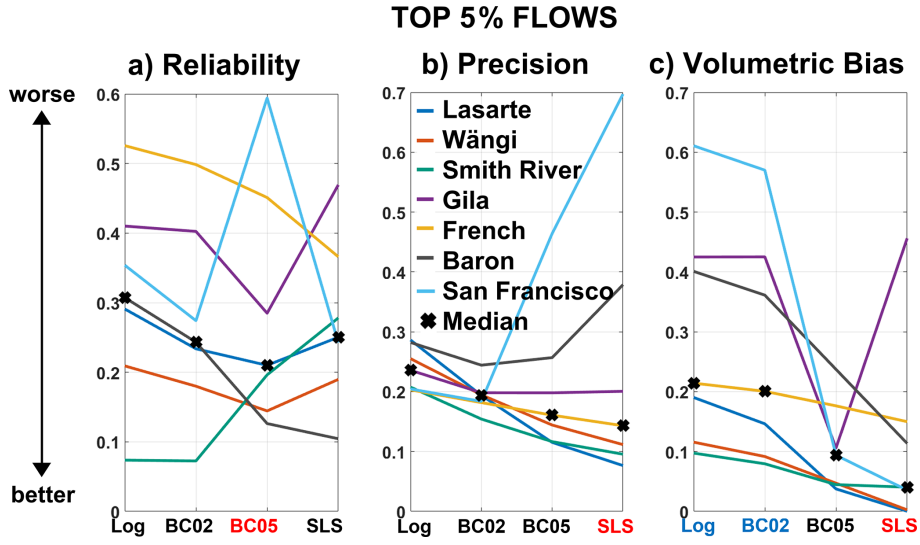


Figure 3. Performance of residual error models across all catchments for top 5 % of flows. Reliability is comparable for the transformations. The median precision and volumetric bias improve monotonically from Log to SLS, but these changes are not statistically significant for Precision. The best-median transformation is indicated in red. The transformations that are not statistically significantly better than the best-median transformation are shown in black bold and the transformations that are statistically significantly worse are shown in blue.

Table 3. Summary of best-median residual error model and best-metric residual error model across all catchments for top 5 % of flows.

Metric	Best-metric residual error model (best-median residual error model in bold)
Reliability	Log, BC02, **BC05**, SLS
Precision	Log, BC02, BC05, **SLS**
Volumetric Bias	BC05, **SLS**

Table 4. Average reliability and precision across catchments, autoregressive models (row labelled "Ave AR", values in bold) and aggregation time scales (column labelled "Ave (h, d, m)", where h, d, and m denote hourly, daily, and monthly; values are in bold), for the residual error models in this study.

		Reliability				Precision			
		Hourly	Daily	Monthly	Ave (h, d, m)	Hourly	Daily	Monthly	Ave (h, d, m)
Log	AR1	0.080	0.083	0.158	**0.107**	0.356	0.341	0.231	**0.309**
	AR2	0.078	0.079	0.216	**0.124**	0.351	0.326	0.158	**0.278**
	AR3	0.078	0.080	0.207	**0.122**	0.351	0.328	0.164	**0.281**
	Ave AR	**0.079**	**0.081**	**0.193**	–	**0.352**	**0.332**	**0.184**	–
BC02	AR1	0.105	0.104	0.151	**0.120**	0.368	0.352	0.228	**0.316**
	AR2	0.106	0.099	0.208	**0.138**	0.368	0.337	0.143	**0.283**
	AR3	0.106	0.100	0.191	**0.133**	0.368	0.340	0.152	**0.287**
	Ave AR	**0.106**	**0.101**	**0.183**	–	**0.368**	**0.343**	**0.174**	–
BC05	AR1	0.170	0.166	0.217	**0.184**	0.506	0.480	0.279	**0.422**
	AR2	0.174	0.160	0.292	**0.208**	0.512	0.447	0.148	**0.369**
	AR3	0.174	0.163	0.262	**0.200**	0.512	0.455	0.170	**0.379**
	Ave AR	**0.173**	**0.163**	**0.257**	–	**0.510**	**0.461**	**0.199**	–
SLS	AR1	0.272	0.260	0.333	**0.288**	1.063	0.956	0.400	**0.806**
	AR2	0.276	0.264	0.385	**0.308**	1.079	0.790	0.183	**0.684**
	AR3	0.276	0.262	0.356	**0.298**	1.079	0.848	0.223	**0.717**
	Ave AR	**0.275**	**0.262**	**0.358**	–	**1.074**	**0.864**	**0.269**	–

5 Discussion

5.1 Performance of probabilistic predictions for all flows

When considering performance across all flows, our results clearly favour the Log transformation: this transformation provides the single best-metric performance in terms of reliability, it is one of two best-metric transformations in terms of precision, and it is statistically no worse than other transformations in terms of volumetric bias. The BC02 transformation is seen to be the next best transformation – while it has worse reliability than the Log transformation, it achieves the best-median performance in terms of precision and is statistically no worse than other transformations for bias. McInerney et al. (2017) reported similar findings at the daily scale for reliability and precision. These rankings are generally consistent across the seven case study catchments, although for bias there is greater variability.

When considering the best-metric transformations, we find that increasing λ from 0 to 0.2, i.e., from Log to BC02, tends to improve precision at the expense of reliability. Conversely, reliability improves monotonically as λ is reduced, with the best reliability achieved by the Log transformation. This finding is also similar to McInerney et al. (2017). The consistency of findings at the hourly and daily scales is reassuring and can be expected to facilitate the selection of residual error models for practical work.

In practice, a precise prediction is desirable only to the extent that it is also reliable. Notably, a precise but unreliable prediction is overconfident, and its predictive uncertainty estimates can be misleading (Klotz et al., 2022; Li et al., 2021). McInerney et al. (2017) report better precision for intermediate values of λ (BC02 and BC05), whereas we find that at hourly time scales, Log is not statistically significantly worse than BC02 in terms of precision. For example, compare our Fig. 1 versus Fig. 2 in McInerney et al. (2017).

For volumetric bias (Fig. 1c), there are some differences and some similarities with McInerney et al. (2017). BC05 is a best-metric transformation in both studies, but SLS performs better in this study than in McInerney et al. (2017). In our hourly study SLS achieves the lowest volumetric bias, although the differences among schemes are not statistically significant. In contrast, in the daily study of McInerney et al. (2017), BC05 achieved the lowest volumetric bias, whereas

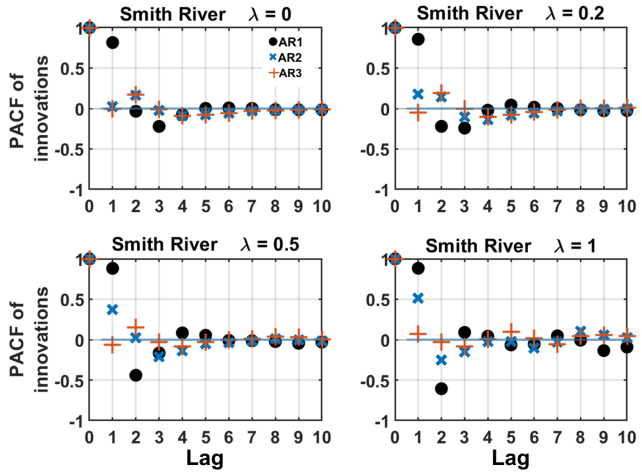


Figure 4. Comparison of AR models in the Smith River catchment. Each panel shows the partial autocorrelation function (PACF) of innovations estimated when the AR1, AR2 and AR3 models are used in the residual error model. The four panels report the analysis for the four values of the Box-Cox transformation parameter λ considered in this study. PACF values close to zero indicate that the error autocorrelation at these lags is well represented by the AR model (see Sect. 3.6.3). In this catchment, the AR3 model provides the best representation of the autocorrelation of residual errors (greatest reduction in the partial autocorrelation of innovations), for all values of λ.

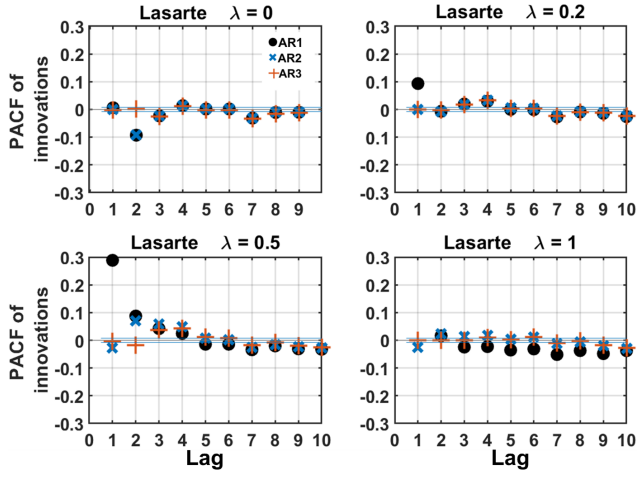


Figure 5. Comparison of AR models – Illustration 2. Lasarte catchment. Here we see a different behavior than in the Smith River catchment, with AR1–AR3 models all providing comparable reduction in PACF except when $\lambda = 0.5$.

SLS exhibited substantially larger bias and ranked among the poorest-performing schemes.

The stabilisation of streamflow variance by the Log transformation avoids the objective function being dominated by high flows, and thus provides more balanced performance across the entire flow range. A similar finding was reported

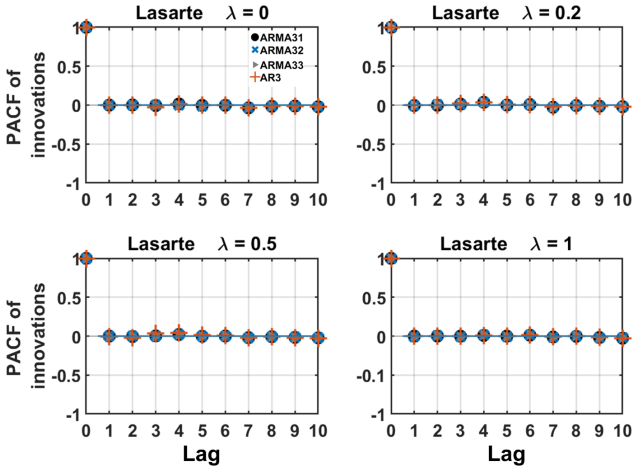


Figure 6. Representative behavior resulting from including MA terms in the autoregressive model, illustrated using the PACF of innovations in the Lasarte catchment. Here, MA terms provide little additional benefit beyond the AR model.

by McInerney et al. (2017), and indeed represents the underlying rationale for using variance-stabilising transformations.

The favourable performance of the Log transformation is consistent with the residuals being skewed and heavy tailed – as already reported at the daily scale by McInerney et al. (2017). The Countable Mixtures of Asymmetric Laplacians (CMAL) method (Klotz et al., 2022) is another model that generates asymmetric heavy tails, which likely explains its good performance in representing streamflow uncertainty.

Note that in principle it is possible to infer λ alongside other parameters. However, as shown by McInerney et al. (2017), the parameter inference becomes dominated by low flows, which pushes λ into negative values and in turn produces extremely heavy tails and unstable predictions, especially for high flows. Fixing the value of λ a priori is a pragmatic workaround to this problem.

5.2 Performance of probabilistic predictions for top 5 % of flows

For the top 5 % of the flows, the BC05 transformation achieves the best-median performance for reliability and SLS has the best-median performance for precision and volumetric bias. However, for reliability and precision none of the transformations are statistically significantly worse than the others, whereas for volumetric bias the BC05 transformation is not statistically significantly worse than SLS (although Log and BC02 are significantly worse). These findings are important because they suggest that residual error models based on transformations recommended for the full range of flows are not significantly disadvantaged when applied to high flows, at least in terms of reliability and precision. The increased variability in catchment-by-catchment perfor-

mance for the top 5 % of flows compared to that for all flows is also noted.

Finally, note that "high" flows could be defined using criteria other than our 5 % threshold. Our purpose here is to explore the performance of streamflow transformations under a commonly used definition of high flows. The top 5 % of flows is a widely used threshold for high flows; for example, see standardized datasets such as CAMELS or EStreams (Addor et al., 2017; do Nascimento et al., 2024). See also Westerberg and McMillan (2015) for other high-flow percentiles. Alternative classification methods have also been employed, e.g. variance-based methods (e.g., see Fischer et al., 2021).

5.3 Aggregated results

The aggregation from hourly to daily and further to monthly is examined as a test for the "seamlessness" of the predictions, as well as an indirect test of the treatment of autocorrelation.

Hourly to daily aggregation has little effect on performance, as reflected in the reliability and precision metrics shown in Table 4. This behavior is beneficial in operational contexts, because it implies that the probabilistic model – here based on the Box-Cox transformation and ARMA model – does not require re-calibration at the longer time scale. This finding can be related to the work of McInerney et al. (2022) who found that aggregating the predictions of the Multi-Temporal Hydrological Residual Error (MuTHRE) model from daily to monthly time scales did not result in a loss of reliability. We consider it a worthwhile advance towards "seamless" predictions using a single hourly product.

The finding that hourly to daily aggregation largely preserves performance also suggests that the autocorrelation of residual errors is represented sufficiently well across these two time scales. It was shown in previous work that poor representation of error autocorrelation leads to poor reliability of aggregated probabilistic predictions (Evin et al., 2014). For example, if error autocorrelation is ignored altogether, the uncorrelated errors in the predictions cancel out during the aggregation, resulting in grossly under-estimated uncertainty. For this reason, seamless prediction methods in the literature have paid considerable attention to the representation of error autocorrelation – e.g., McInerney et al. (2020) included a representation of dynamic biases with memory of up to 30 d to enable reliable aggregation of probabilistic predictions from daily to monthly scales. Our analysis suggests the AR3 model is sufficient at least across the hourly to daily time scales.

Further aggregation to monthly scale results in deterioration in reliability, but improvement in precision, in most catchments. We speculate that the loss of reliability is due mainly to deficiencies in the treatment of longer scale error autocorrelation and more general error persistence patterns – the changes in processes affecting aggregation from daily to monthly scales are more substantial than changes affecting aggregation from hourly to daily scales. In addition, the monthly scale analysis by necessity has fewer data points, which is likely to result in noisier estimates of the performance metrics and affect the comparisons with metrics at daily and hourly scales. These questions could be explored using the longer-term persistence components included in the MuTHRE model of McInerney et al. (2022), as well as using longer data sets where available.

5.4 Practical recommendations for selecting streamflow transformations at hourly scale

In practical applications, the selection of a residual error model is governed by two key considerations: (1) performance in specific metrics of interest, e.g., reliability, precision or bias; and (2) resources, e.g., time, cost and ease of implementation.

Based on the empirical results in this study, we make the following recommendations for hourly predictions across all flow ranges:

a. If reliability is the highest priority: we recommend the Log transformation. This recommendation is similar to the recommendation of McInerney et al. (2017) at daily scale.

b. If precision is the highest priority: we still recommend the Log transformation. The BC02 transformation is also a good option – it is not (statistically significantly) worse than Log for precision while being better in terms of bias. This recommendation is similar to McInerney et al. (2017), who found that the BC02 transformation was the best for precision, even though its other metrics were statistically significantly worse.

c. If low volumetric bias is the highest priority: we recommend the Log or BC02 transformations. Note that McInerney et al. (2017) found that the BC05 transformation was the best-performing transformation for volumetric bias, but (unlike in this study) the other transformations performed significantly worse.

When considering performance stratified to top 5 % of flows, our recommendation remains to use the Log transformation for reliability and precision, with the BC05 transformation recommended if low volumetric bias is of particular importance. Here, it is notable that the transformations recommended earlier for the full flow range are not substantially disadvantaged.

In terms of ease of implementation, all transformations used in this work are straightforward to implement, which is clearly appealing in practical contexts.

Overall, this hourly scale study reaches broadly similar recommendations to the earlier recommendations of McInerney et al. (2017) for daily time scales, notably in terms of recommending the Log and BC02 transformations, with some differences noted above.

5.5 Limitations and avenues for future research

5.5.1 Constant autocorrelation

The change in the hydrograph as recession begins is likely to violate the assumption of constant autocorrelation in the residuals during the rising and falling limbs. Generally, we expect residuals in the falling limb to be more autocorrelated than in the rising limb. Therefore, the (constant) autoregressive model is more suited for representing uncertainty in the falling limb.

Previous work on representing this behaviour in residual error models include Li et al. (2017) where separate autocorrelation coefficients were used for the rising and falling limbs of the hydrographs, and Ammann et al. (2019) where the autocorrelation parameter was made dependent on rainfall. Note that Li et al. (2017) used a post-processor approach and Ammann et al. (2019) used joint inference. These approaches are recommended for investigation in future work at the hourly scale.

5.5.2 Generality of findings: Other catchments and performance metrics

The study considers 7 catchments, which is relatively fewer catchments than in similar studies at the daily scale, (e.g., Hunter et al., 2021; Li et al., 2016; Li et al., 2017; McInerney et al., 2017). Therefore, the findings are necessarily conditional on the choice of these catchments. For this reason, we recommend further studies with a larger number of catchments. The growing availability of hourly data can help in this endeavour. A natural next step would be to explore hourly scale probabilistic prediction for catchments with different hydroclimatology, such as snow-dominated, ephemeral, arid, and semi-arid. In these catchments, reliability might not be necessarily the priority, e.g., volumetric bias may be more important for water planning purposes in semi-arid catchments.

In addition, the study employs an evaluation approach based on statistical performance metrics (namely reliability, precision and volumetric bias), applied to either the entire hydrograph or the top 5 % of flows. These performance metrics have been widely used in the hydrological literature when assessing the performance of probabilistic models, both in prediction (McInerney et al., 2019; McInerney et al., 2017) as well as forecasting (Huang and Zhao, 2022).

However, while these metrics provide theoretically rigorous assessments of general statistical performance and are therefore appropriate for a general scientific evaluation, practical applications often require more specific performance criteria. For example, a flood forecasting service will logically prioritise performance for high flows, and may do so in the context of true vs false alarms rather than in a general distributional sense. Conversely, drought prediction will logically focus on overall volumes in the context of water supply risk. Still other applications may focus on low-flow thresholds to maintain ecological health of streams, and so forth.

These considerations have been explored in McInerney et al. (2024), which highlighted the importance of probabilistic predictions when undertaking risk assessment. Nevertheless, the optimal choice of probabilistic model – as well as of the deterministic model! – will likely depend on the specific application context.

For this reason, the current study should be seen as a step towards establishing probabilistic models for hourly predictions, which provides a general evaluation of practical methods rather than final recommendations for specific modelling contexts. Uncertainty quantification for hourly streamflow predictions represents an important direction for future work, and in our opinion should include the relatively simple methods (Box-Cox transformations and ARMA models) considered in this study.

6 Conclusions

This study explores hourly scale streamflow predictions and associated uncertainty estimation, using comparatively simple residual error modelling approaches. Empirical performance is reported for seven catchments with humid to semi-arid conditions.

It is found that the choice of streamflow transformation employed in the residual error model has a major impact on predictive performance. When considering performance across all flows, the Log transformation achieves the best reliability of predictions, with the improvement over other transformations being statistically significant of predictions. Moreover, its precision and volumetric bias are not statistically significantly worse than those achieved by Box-Cox transformations with other values of the power parameter (0.2–1). The findings are generally consistent across catchments. When stratifying performance evaluation to the top 5 % of flows, performance across catchments becomes notably more variable. Nevertheless, the Log transformation remains competitive, achieving reliability and precision that are not statistically significantly worse than for the other transformations. Additionally, the BC05 transformation becomes attractive for this stratification, as it achieves lower volumetric bias. These findings are broadly similar to the earlier work of McInerney et al. (2017) with daily data, which also favored residual error models based on the Log, BC02 and BC05 transformations.

The representation of autocorrelation in residual errors is another important consideration, which in this work is investigated using autoregressive models. Comparing AR1, AR2 and AR3 models using PACFs, the autocorrelation is captured much better by the AR2 model than by the AR1 model. In some catchments, the AR3 model captures the autocorrelation better than the AR2 model. Adding moving-average

(MA) terms to the AR3 terms does not provide an improvement.

Finally, the ability to achieve "seamless" aggregation from hourly to longer scales is attractive in practical contexts. When using the Log transformation, model performance when aggregating from hourly to daily is relatively stable and preserves reliability and precision. However, further aggregation to monthly scale suffers from a loss of reliability in most catchments.

These findings provide practical guidance for hourly scale studies in predictive uncertainty quantification. Further work is recommended using a wider range of catchments, models, and performance metrics.

Appendix A: Definitions of performance metrics

The reliability metric is defined as

$$\text{Reliability}\left[\mathbf{Q}, \tilde{\mathbf{q}}\right] = \frac{2}{N_t}\sum_{t=1}^{N_t}\left|F_{\Omega}\left[F_{Q_t}\left(\tilde{q}_t\right)\right] - F_{\text{U}}\left[F_{Q_t}\left(\tilde{q}_t\right)\right]\right| \tag{A1}$$

where F_{U} is the cumulative distribution function (CDF) of the standard uniform distribution, F_{Ω} is the empirical CDF of the set $\Omega = \left\{F_{Q_t}(\tilde{q}_t); t = 1, \ldots, N_t\right\}$, F_{Q_t} is the cumulative distribution function (CDF) of the predictive distribution at time t, and $\tilde{q}_t$ is the corresponding observation at time t. Reliable probabilistic predictions are characterized by the values in the set Ω following a standard uniform distribution, and the reliability metric in Eq. (A1) quantifies this discrepancy. Lower values of the metric indicate better reliabiliy. See Renard et al. (2010) and Evin et al. (2014) for further details.

The precision metric is defined as

$$\text{Precision}\left[\mathbf{q}^{\text{pred}}, \tilde{\mathbf{q}}\right] = \frac{\frac{1}{N_t}\sum_{t=1}^{N_t}\text{SD}\left(\boldsymbol{q}_t^{\text{pred}}\right)}{\frac{1}{N_t}\sum_{t=1}^{N_t}\tilde{q}_t} \tag{A2}$$

where $\text{SD}\left(\mathbf{q}_t^{\text{pred}}\right)$ is the standard deviation of the predictive replicates $\left\{\boldsymbol{q}_t^{\text{pred}(r)}; r = 1, \ldots, N_r\right\}$. The metric represents the average spread of the predictive distributions scaled by the mean observed streamflow. Lower values indicate better precision. See McInerney et al. (2017) for further details.

The volumetric bias metric is defined as

$$\text{Volumetric Bias}\left[\mathbf{q}^{\text{pred}}, \tilde{\mathbf{q}}\right] = \left|\frac{\sum_{t=1}^{N_t}\overline{\boldsymbol{q}}_t^{\text{pred}} - \sum_{t=1}^{N_t}\tilde{q}_t}{\sum_{t=1}^{N_t}\tilde{q}_t}\right| \tag{A3}$$

where $\overline{\mathbf{q}}_t^{\text{pred}}$ is the mean of the predictive replicates at time step t,

$$\overline{\boldsymbol{q}}_t^{\text{pred}} = \frac{1}{N_r}\sum_{t=1}^{N_t}\boldsymbol{q}_t^{\text{pred}(r)} \tag{A4}$$

The numerator in Eq. (A3) represents the difference between the total predicted and observed volumes, and the denominator scales this difference by the total observed volume. Lower values of this metric indicate lower (better) volumetric bias. See McInerney et al. (2017) and McInerney et al. (2018) for further details.

Appendix B: Performance of residual error models across catchments, models and time scales

This Appendix details the calculation of average reliability and precision metrics reported in Table 4.

Let $\text{Reliability}_{c,N_\phi,\tau}^{(\lambda)}$ denote the reliability metric of a residual error model with Box-Cox power parameter λ and AR model of order N_ϕ, applied in catchment c at time scale τ. Here, $\tau =$ h, d, m denotes, respectively, hourly, daily and monthly time scales.

Average reliability across catchments is computed as

$$\overline{\text{Reliability}}_{N_\phi,\tau}^{(\lambda)} = \frac{1}{N_c}\sum_{c=1}^{N_c}\text{Reliability}_{c,N_\phi,\tau}^{(\lambda)} \tag{B1}$$

where $N_c = 7$ is the number of catchments.

Additional averaging across AR models is implemented as

$$\overline{\overline{\text{Reliability}}}_{\tau}^{(\lambda)} = \frac{1}{N_{\text{AR}}}\sum_{i=1}^{N_{\text{AR}}}\overline{\text{Reliability}}_{N_\phi^{(i)},\tau}^{(\lambda)} \tag{B2}$$

where $N_\phi^{(i)}$ is the order of the ith AR model and $N_{\text{AR}} = 3$ is the number of AR models.

Alternatively, additional averaging across time scales is implemented as

$$\overline{\overline{\text{Reliability}}}_{N_\phi}^{(\lambda)} = \frac{1}{N_\tau}\sum_{\tau\in\{\text{h, d, m}\}}\overline{\text{Reliability}}_{N_\phi,\tau}^{(\lambda)} \tag{B3}$$

where $N_\tau = 3$ is the number of time scales.

The same averaging procedures are applied to the precision metric.

Notably, the performance metrics are dimensionless, which avoids dependence on scale-specific flow ranges, etc.

Data availability. The study data are deposited in https://doi.org/10.5281/zenodo.18379721 (Prieto et al., 2026).

Author contributions. Research question: CP, DK, and FF designed the research question. Data: CP, JK, and FF prepared the data. Methodology, model runs, predictions, and performance evaluation: CP, DK, and FF, with support from DM and MT. Writing: CP wrote the first draft of the manuscript. DK, FF, and JK contributed substantially to revisions. DM and MT critically reviewed the manuscript. CA reviewed the manuscript.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Hydrology and Earth System Sciences*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. The lead author acknowledges institutional support from IHCantabria and the University of Cantabria, as well as support as an academic guest at ETH Zurich and as a research visitor at Eawag, where part of this work was carried out. The authors thank Andreas Scheidegger and Grey Nearing for insightful discussions on uncertainty at the hourly time scale.

Financial support. This research has been supported by ETH Zurich during the lead author's academic guest appointment, the Government of Cantabria through the FÉNIX programme (grant no. 2020.03.03.322B.742.09), and the MCIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+) through the Ramón y Cajal programme (grant no. RYC2024-050770-I).

Review statement. This paper was edited by Christa Kelleher and reviewed by two anonymous referees.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Ammann, L., Fenicia, F., and Reichert, P.: A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation, Hydrol. Earth Syst. Sci., 23, 2147–2172, https://doi.org/10.5194/hess-23-2147-2019, 2019.

Arora, V. K.: The use of the aridity index to assess climate change effect on annual runoff, J. Hydrol., 265, 164–177, https://doi.org/10.1016/S0022-1694(02)00101-4, 2002.

Ayzel, G. and Heistermann, M.: The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: A case study for six basins from the CAMELS dataset, Comput. Geosci., 149, 104708, https://doi.org/10.1016/j.cageo.2021.104708, 2021.

Basque Water Agency (URA): Gauging station data and data-request service, https://www.uragentzia.euskadi.eus/datos-de-estaciones-de-aforo/webura00-contents/es/, last access: 15 July 2026.

Bauer, D. F.: Constructing Confidence Sets Using Rank Statistics, J. Am. Stat. Assoc., 67, 687–690, https://doi.org/10.1080/01621459.1972.10481279, 1972.

Bieroza, M., Acharya, S., Benisch, J., Ter Borg, R. N., Hallberg, L., Negri, C., Pruitt, A., Pucher, M., Saavedra, F., Staniszewska, K., Van't Veen, S. G. M., Vincent, A., Winter, C., Basu, N. B., Jarvie, H. P., and Kirchner, J. W.: Advances in Catchment Science, Hydrochemistry, and Aquatic Ecology Enabled by High-Frequency Water Quality Measurements, Environ. Sci. Technol., 57, 4701–4719, https://doi.org/10.1021/acs.est.2c07798, 2023.

Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Čanjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnová, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Živković, N.: Changing climate both increases and decreases European river floods, Nature, 573, https://doi.org/10.1038/s41586-019-1495-6, 2019.

Box, G. E. P. and Cox, D. R.: An Analysis of Transformations, J. R. Stat. Soc. Ser. B Methodol., 26, 211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x, 1964.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C.: Time Series Analysis: Forecasting and Control, Prentice Hall, ISBN 978-0470272848, 1994.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M.: Time series analysis: forecasting and control, John Wiley and Sons, ISBN: 978-1-118-67502-1, 2015.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, https://doi.org/10.1029/2007wr006735, 2008.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resour. Res., 48, https://doi.org/10.1029/2011WR011721, 2012.

Dal Molin, M., Schirmer, M., Zappa, M., and Fenicia, F.: Understanding dominant controls on streamflow spatial variability to set up a semi-distributed hydrological model: the case study of the Thur catchment, Hydrol. Earth Syst. Sci., 24, 1319–1345, https://doi.org/10.5194/hess-24-1319-2020, 2020.

Dal Molin, M., Kavetski, D., Albert, C., and Fenicia, F.: Exploring Signature-Based Model Calibration for Streamflow Prediction in Ungauged Basins, Water Resour. Res., 59, e2022WR031929, https://doi.org/10.1029/2022WR031929, 2023.

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., Savenije, H., Thirel, G., and Willems, P.: Looking beyond general metrics for model comparison – lessons from an international model intercomparison study, Hydrol. Earth Syst. Sci., 21, 423–440, https://doi.org/10.5194/hess-21-423-2017, 2017.

Dieter, F. H. and Arns, S. A.: Estimating Instantaneous Peak Flow from Mean Daily Flow Data, J. Hydrol. Eng., 8, https://doi.org/10.1061/(ASCE)1084-0699(2003)8:6(365), 2003.

do Nascimento, T. V. M., Rudlang, J., Höge, M., van der Ent, R., Chappon, M., Seibert, J., Hrachowitz, M., and Fenicia, F.: EStreams: An integrated dataset and catalogue of streamflow, hydro-climatic and landscape variables for Europe, Sci. Data, 11, 879, https://doi.org/10.1038/s41597-024-03706-1, 2024.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J. Hydrol., 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.

Evin, G., Kavetski, D., Thyer, M., and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour. Res., 49, 4518–4524, https://doi.org/10.1002/wrcr.20284, 2013.

Evin, G., Thyer, M., Kavetski, D., McInerney, D., and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resour. Res., 50, 2350–2375, https://doi.org/10.1002/2013wr014185, 2014.

Fischer, S., Schumann, A., and Bühler, P.: A statistics-based automated flood event separation, Journal of Hydrology X, 10, 100070, https://doi.org/10.1016/j.hydroa.2020.100070, 2021.

FOEN: Federal Office for the Environment, https://www.bafu.admin.ch/bafu/en/home.html, last access: 21 July 2026.

Forte, G., De Falco, M., Santo, A., Gautam, D., and Santangelo, N.: Flash flood impacts and vulnerability mapping at catchment scale: Insights from southern Apennines, Eng. Geol., 350, 107988, https://doi.org/10.1016/j.enggeo.2025.107988, 2025.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci., 25, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021, 2021.

Hamilton, J. D.: Time Series Analysis, Princeton University Press, https://doi.org/10.2307/j.ctv14jx6sm, 1994.

Huang, Z. and Zhao, T.: Predictive performance of ensemble hydroclimatic forecasts: Verification metrics, diagnostic plots and forecast attributes, WIREs Water, 9, e1580, https://doi.org/10.1002/wat2.1580, 2022.

Hunter, J., Thyer, M., McInerney, D., and Kavetski, D.: Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions, J. Hydrol., 603, https://doi.org/10.1016/j.jhydrol.2021.126578, 2021.

Kavetski, D., Fenicia, F., and Clark, M. P.: Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, Water Resour. Res., 47, https://doi.org/10.1029/2010WR009525, 2011.

Kirchner, J. W.: A double paradox in catchment hydrology and geochemistry, Hydrol. Proc., 17, 871–874, https://doi.org/10.1002/hyp.5108, 2003.

Kirchner, J. W., Feng, X., Neal, C., and Robson, A. J.: The fine structure of water-quality dynamics: the (high-frequency) wave of the future, Hydrol. Proc., 18, 1353–1359, https://doi.org/10.1002/hyp.5537, 2004.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673-1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Koutsoyiannis, D. and Montanari, A.: Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions, Water Resour. Res., 58, https://doi.org/10.1029/2021wr031215, 2022.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, https://doi.org/10.5194/hess-11-1267-2007, 2007.

Li, M., Robertson, D. E., Wang, Q. J., Bennett, J. C., and Perraud, J.-M.: Reliable hourly streamflow forecasting with emphasis on ephemeral rivers, J. Hydrol., 598, https://doi.org/10.1016/j.jhydrol.2020.125739, 2021.

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

Li, M., Wang, Q. J., Robertson, D. E., and Bennett, J. C.: Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs, J. Hydrol., 555, https://doi.org/10.1016/j.jhydrol.2017.10.057, 2017.

Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants, Doctorat spécialité Sciences de l'eau, ENGREF Paris, Sde Irstea Gip-be Inrae Hycar, https://hal.inrae.fr/tel-02587642 (last access: 25 August 2026), 2005.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resour. Res., 53, 2199–2239, https://doi.org/10.1002/2016wr019168, 2017.

McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., and Kuczera, G.: A simplified approach to produce probabilistic hydrological model predictions, Environ. Modell. Softw., 109, 306–314, https://doi.org/10.1016/j.envsoft.2018.07.001, 2018.

McInerney, D., Kavetski, D., Thyer, M., Lerat, J., and Kuczera, G.: Benefits of Explicit Treatment of Zero Flows in Probabilistic Hydrological Modeling of Ephemeral Catchments, Water Resour. Res., 55, 11035–11060, https://doi.org/10.1029/2018wr024148, 2019.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., and Kuczera, G.: Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting, Water Resour. Res., 56, https://doi.org/10.1029/2019wr026979, 2020.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., and Kuczera, G.: Improving the Reliability of Sub-Seasonal Forecasts of High and Low Flows by Using a Flow-Dependent Nonparametric Model, Water Resour. Res., 57, https://doi.org/10.1029/2020wr029317, 2021.

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., and Kuczera, G.: Seamless streamflow forecasting at daily to monthly scales: MuTHRE lets

you have your cake and eat it too, Hydrol. Earth Syst. Sci., 26, 5669–5683, https://doi.org/10.5194/hess-26-5669-2022, 2022.

McInerney, D., Thyer, M., Kavetski, D., Westra, S., Maier, H. R., Shanafield, M., Croke, B., Gupta, H., Bennett, B., and Leonard, M.: Neglecting hydrological errors can severely impact predictions of water resource system performance, J. Hydrol., 634, 130853, https://doi.org/10.1016/j.jhydrol.2024.130853, 2024.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, J. Hydrol., 400, 83–94, https://doi.org/10.1016/j.jhydrol.2011.01.026, 2011.

Montanari, A. and Koutsoyiannis, D.: Uncertainty estimation for environmental multimodel predictions: The BLUECAT approach and software, Environ. Modell. Softw., 188, 106419, https://doi.org/10.1016/j.envsoft.2025.106419, 2025.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, Hydrol. Earth Syst. Sci., 26, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022, 2022.

Paltan, H., Waliser, D., Lim, W. H., Guan, B., Yamazaki, D., Pant, R., and Dadson, S.: Global Floods and Water Availability Driven by Atmospheric Rivers, Geophys. Res. Lett., 44, 10387–310395, https://doi.org/10.1002/2017GL074882, 2017.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Prieto, C., Patel, D., and Han, D.: Preface: Advances in flood risk assessment and management, Nat. Hazards Earth Syst. Sci., 20, 1045–1048, https://doi.org/10.5194/nhess-20-1045-2020, 2020.

Prieto, C., Kavetski, D., Le Vine, N., Álvarez, C., and Medina, R.: Identification of Dominant Hydrological Mechanisms Using Bayesian Inference, Multiple Statistical Hypothesis Testing, and Flexible Models, Water Resour. Res., 57, e2020WR028338, https://doi.org/10.1029/2020WR028338, 2021.

Prieto, C., Patel, D., Han, D., Dewals, B., Bray, M., and Molinari, D.: Preface: Advances in pluvial and fluvial flood forecasting and assessment and flood risk management, Nat. Hazards Earth Syst. Sci., 24, 3381–3386, https://doi.org/10.5194/nhess-24-3381-2024, 2024.

Prieto, C., Kavetski, D., Fenicia, F., Kirchner, J., McInerney, D., Thyer, M., and Álvarez, C.: Data: Simple Box-Cox probabilistic models for hourly streamflow predictions, Zenodo [data set], https://doi.org/10.5281/zenodo.18379721, 2026.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, https://doi.org/10.1029/2009wr008328, 2010.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S. W.: Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, Water Resour. Res., 47, https://doi.org/10.1029/2011wr010643, 2011.

Romero-Cuellar, J., Arabzadeh, R., Craig, J. R., Tolson, B. A., and Mai, J. A multi-model evaluation of probabilistic streamflow predictions via residual error modelling, J. Hydrol., 635, 131152, https://doi.org/10.1016/j.jhydrol.2024.131152, 2024.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46, https://doi.org/10.1029/2009wr008933, 2010.

Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, Water Resour. Res., 16, 430–442, https://doi.org/10.1029/WR016i002p00430, 1980.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, https://doi.org/10.1029/2008wr006825, 2009.

United States Environmental Protection Agency: An Approach for Using Load Duration Curves in the Development of TMDLs (EPA 841-B-07-006), https://www.epa.gov/tmdl/approach-using-load-duration-curves-development-tmdls (last access: 25 August 2026), 2007.

van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., and Lobligeois, F.: The influence of conceptual model structure on model performance: a comparative study for 237 French catchments, Hydrol. Earth Syst. Sci., 17, 4227–4239, https://doi.org/10.5194/hess-17-4227-2013, 2013.

Vrugt, J. A.: Distribution-Based Model Evaluation and Diagnostics: Elicitability, Propriety, and Scoring Rules for Hydrograph Functionals, Water Resour. Res., 60, e2023WR036710, https://doi.org/10.1029/2023WR036710, 2024.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resour. Res., 48, https://doi.org/10.1029/2011WR010973, 2012.

Wani, O., Scheidegger, A., Cecinati, F., Espadas, G., and Rieckermann, J.: Exploring a copula-based alternative to additive error models – for non-negative and autocorrelated time series in hydrology, J. Hydrol., 575, 1031–1040, https://doi.org/10.1016/j.jhydrol.2019.06.006, 2019.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951–3968, https://doi.org/10.5194/hess-19-3951-2015, 2015.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Rev. Geophys., 52, 522–555, https://doi.org/10.1002/2014RG000464, 2014.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, https://doi.org/10.1029/2007WR006716, 2008.

Yin, J., Gao, Y., Chen, R., Yu, D., Wilby, R., Wright, N., Ge, Y., Bricker, J., Gong, H., and Guan, M.: Flash floods: why are more of them devastating the world's driest regions?, Nature, 615, 212–215, https://doi.org/10.1038/d41586-023-00626-9, 2023.

Zheng, Y., Coxon, G., Woods, R., Li, J., and Feng, P.: Controls on the Spatial and Temporal Patterns of Rainfall-Runoff Event Characteristics – A Large Sample of Catchments Across Great Britain, Water Resour. Res., 59, e2022WR033226, https://doi.org/10.1029/2022WR033226, 2023.