https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Simple Box-Cox probabilistic models for hourly streamflow predictions

Cristina Prieto[1,2,3], Dmitri Kavetski[4,1], Fabrizio Fenicia[3], James 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, Environmental and Mining Engineering, University of Adelaide, Adelaide, SA, Australia
[5]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[6]Department of Earth and Planetary Science, University of California, Berkeley, California, USA

Correspondence to: Cristina Prieto (prietoc@unican.es)

Abstract. The increasing availability of hourly scale hydrological data offers valuable benefits for advancing our scientific understanding of catchment processes and improving operational forecasting capabilities. This work contributes to streamflow predictions at the hourly scale by investigating practical methods for uncertainty quantification using probabilistic predictions. We examine common approaches for representing the heteroscedasticity of streamflow errors using the Box-Cox (BC) transformation and common approaches for representing the persistence of streamflow errors using auto-regressive (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 the BC02 and BC05 transformations respectively. The results also tend to favor the AR2 and AR3 models over the AR1 model in representing persistence of errors, with the addition of moving average terms providing little additional benefit. The study findings are broadly consistent with earlier work with daily data, and provide practical guidance for hourly scale studies in predictive uncertainty quantification that is accessible to a wide range of hydrologists. 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 forecasting 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 forecasting tools.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Representations of wetting and drying dynamics benefit greatly from hourly data to calibrate and evaluate model behavior. For example, response to precipitation is highly sensitive to antecedent wetness conditions (Kirchner, 2003; Zheng et al., 2023; Westerberg and Mcmillan, 2015), 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 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 mass balance considerations, whereas hourly predictions require resolving processes such as overland flow, channel routing, and interception dynamics. These processes introduce added complexity and typically demand more detailed model structures and data (see, for example, (Bieroza et al., 2023; Kirchner et al., 2004)). From a more applied perspective, peak flows may last only a few hours, so forecasts of daily flows can greatly underestimate flood peaks (Fill Heinz and Steiner Alexandre, 2003; Bartens et al., 2024). Hourly scales are particularly important in small and mesoscale catchments (less than about 10,000 km2) due to faster response, e.g., causing “flash floods” (e.g. Forte et al. (2025)).

This work contributes to streamflow predictions at the hourly scale, in the context of uncertainty quantification. We focus on practical methods for producing probabilistic predictions, in order to provide a characterization of predictive uncertainty that is accessible to a wide range of hydrologists. The pragmatic focus on (relatively) simple approaches is intended to overcome an arguably common reticence among practitioners to use probabilistic modeling 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., Esse et al. 2013; de Boer-Euseret al. 2017; Li et al. 2017] 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 Term 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 modeling approaches, including machine learning, physically based modeling, and conceptual modeling. Here we confine our analysis to conceptual modeling 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., Sorooshian and Dracup, 1980; McInerney et al. 2017], 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 [Wani et al. 2019, (Li et al., 2016; Mcinerney et al., 2018;

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Mcinerney et al., 2017)]. Reliable uncertainty quantification provides important practical benefits, e.g. (Mcinerney et al., 2024) demonstrated that neglecting hydrological model errors can lead to severe underestimates 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 (Sorooshian & Dracup, 1980; Bates & Campbell, 2001; Evin et al., 2013; Smith et al., 2015; Sun et al., 2017; Amman 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 from a pre-calibrated 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 the joint approach include conceptual models (e.g., Ammann et al. 2019) and LSTM models (Klotz et al [2022]). However, while theoretically more appealing, joint inference can suffer from identifiability problems 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-Cox [1964] transformation with power parameter of 0.2 or 0.5 to reduce heteroscedasticity, followed by the application of a first order autoregressive model (AR1). A 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 the prediction uncertainty is harder to characterize on this time scale.

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 Amman et al. 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)

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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 persistence in hourly residuals has also received attention. The studies by (Li et al., 2021; Li et al., 2016; Li et al., 2017) 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. Amman 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)).

Post-processing methods have shown practical success at daily time scales, both with conceptual models and LSTM models (McInerney et al [2017], Hunter et al. [2021], (Li et al., 2021); (Klotz et al. 2022; 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 transformation for the heteroscedasticity and only a single model for 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 case studies, they have generally been limited to deterministic model analyses (Blöschl et al., 2019; Prieto et al., 2024; Prieto et al., 2020; Addor et al., 2017; Paltan et al., 2017) (Kirchner, 2003; Nevo et al., 2022), with some exceptions being McInerney et al. (2021) where flow stratified performance is reported for probabilistic predictions at the daily scale.

Another practical aspect of hydrological prediction that received recent attention in the literature is the ability to achieve "seamless" prediction, including predictions that remain reliable when aggregated to coarser time scales, e.g. from daily to

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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:

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

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

Aim 3: Explore additional aspects, namely

- Performance of error models for the top 5% of flows (stratified flow performance).
- Performance of error models at time scales aggregated from hourly to daily and monthly.

For 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 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 $q_t^{\theta_h}$ 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),

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

A probabilistic model of streamflow at time t, $Q_t(\boldsymbol{\theta}; \mathbf{x}_{1:t})$, with probability density function (pdf) $p(q_t \,|\, \boldsymbol{\theta}, \mathbf{x}_{1:t})$ 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.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

2.2 Residual error model

2.2.1 Box-Cox transformation to represent error heteroscedasticity

Consider the streamflow distribution

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

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

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

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

where $z(q)$ is a transformation function with parameters $\boldsymbol{\theta}_z$ and η_t is the normalized residual error with parameters $\boldsymbol{\theta}_z$. 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 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_z, \boldsymbol{\theta}_\eta\}$ and includes additional parameters describing the transformation function z and properties of normalized residuals η.

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. Note that $\lambda = 0.2$ was recommended by (Mcinerney et al., 2017) as the most appropriate for daily streamflow predictions.

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

The offset A is used to avoid numerical problems when $q \approx 0$.

2.2.2 ARMA models to represent error autocorrelation

2.2.2.1 AR models

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

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

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

where y_t is the innovation (random component or "noise" term) at time t and $\boldsymbol{\phi} = \{\phi_i ; i = 1,\ldots, N_\phi\}$ 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 $\boldsymbol{\theta}_\eta = \{\sigma_y, \boldsymbol{\phi}\}$.

2.2.2.2 ARMA models

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 parameters so that $\boldsymbol{\theta}_\eta = \{\sigma_y, \boldsymbol{\phi}, \boldsymbol{\varphi}\}$.

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.

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 $\boldsymbol{\theta}_z$,

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

$$\hat{\boldsymbol{\theta}}_h = \underset{\boldsymbol{\theta}_h}{\text{argmax}}\ \Phi_{\text{obj}} \tag{10}$$

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

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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

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

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 unlike (Li et al., 2017) we specify a residual error model with zero mean, see equation (7), because the same transformation is applied to residuals in Stages 1 and 2 (see Hunter et al., 2021). This model assumption was verified empirically by confirming that the mean of the estimated innovations was close to zero.

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(p) persistence structure is generated as follows.

1) Sample innovations from a Gaussian distribution

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

2) Calculate the residuals using

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

3) Apply the inverse transformation

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

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

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

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

where N_r is the number of replicates.

3 Case study

This section described 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 evaluations of the statistical performance of the probabilistic model, which as defined in Equation (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), Wangi (Switzerland), Smith River (CA, USA), Gila (New Mexico, USA), French Broad (NC, USA), Baron (Oklahoma, USA), San Francisco (AZ, USA). These catchments span humid to arid conditions, see Table 1 for details. Hourly rainfall, streamflow and potential evapotranspiration are provided by the Basque Water agency (URA) for Lasarte, BAFU for Wängi and MOPEX for Smith River, Gila, French, Baron and San Francisco catchments. To avoid additional modelling complications, we exclude catchments with ephemeral flow and/or snow-dominated hydrology.

Table 1. Area, mean altitude, aridity (PET/P), rainfall runoff coefficient (Q/P) and Arora [2002] classification for the selected catchments

Catchment	Location	Area, km^2	Mean altitude, m	PET/P	Q/P	Classification, (Arora 2002)
Lasarte	Basque Country, Spain	861	837	0.51	0.52	Humid
Wängi	Wängi, Switzerland	80.15	592.14	0.50	0.60	Humid
Smith	California, USA	614	24.2	0.28	0.73	Humid
Gila	New Mexico, USA	1864	1418.8	2.22	0.09	Semi-arid
French Broad	North Carolina, USA	945	594.4	0.43	0.59	Humid
Baron Fork	Oklahoma, USA	312	213.7	0.99	0.3	Sub-humid

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

San Francisco	Arizona, USA	2763	1047.3	2.33	0.08	Semi-arid

3.2 Deterministic hydrological model

The conceptual hydrological model Génie Rural à 4 paramétres Horaires (GR4H) is employed for the deterministic component in equation (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 Figure 1 of Perrin et al (2003).

GR4H has four calibration parameters: the capacity of the production store θ_1 (mm), the groundwater exchange coefficient θ_2 (mm/h) 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 (hours).

3.3 Residual error model: streamflow transformations to represent heteroscedasticity of residuals

The application of the Box-Cox transformation in this study considers four fixed values of the power parameter: λ = 0 (equivalent to the logarithmic transformation, here denoted "Log"), λ = 1 (no transformation, corresponding to the Simple Least Squares method, denoted "SLS"), λ = 0.5 (square root transformation, here denoted "BC05") and λ = 0.2 (recommended at daily scale in McInerney et al. 2017; denoted "BC02"). The shift parameter is fixed at A = 0.0013 mm/h.

3.4 Model evaluation – split sample validation

A split sample validation approach is employed. In a given catchment, the data is split into two periods. The (probabilistic) model is calibrated on one 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 in the first period. The generated streamflow from the two periods is then concatenated into a single long time series of the length of the full data set. A warmup period is employed, comprising of 2-3 years prior to the calibration period.

3.5 Model evaluation – experiments

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;

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

2. Top 5% of flows, defined as the subset of streamflow values in periods 1 and 2 exceeding 5% of optimized streamflow in periods 2 and 1 respectively – i.e., the threshold in a given validation period is set according to the model streamflow in the associated calibration period. This stratification approach ensures consistency across multiple periods despite potentially different streamflow characteristics. The top 5% of flows is included in the evaluation given that many hydrological applications focus on higher flows (Addor et al., 2017). Alternative definitions of "high" flows as the 2-10% top flows have also been used in the literature (Yilmaz et al., 2008) [USEPA, 2007].

3. Aggregated streamflow time series. We consider aggregations to daily and monthly scales. This performance evaluation is included given the interest is "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).

3.6 Performance metrics and evaluation approach

The probabilistic predictions are evaluated using statistical reliability, precision and bias metrics used extensively in previous studies on probabilistic prediction in hydrology (see, e.g., Renard et al., 2010; Evin et al., 2014; Hunter et al., 2021; and many others), as described in the sections below. The implications of this choice of evaluation metrics are discussed later in Section **¡Error! No se encuentra el origen de la referencia.**.

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

3.6.1 Reliability, Precision and Volumetric Bias metrics

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 Equation 23a and 23b of Renard et al. (2010), which was derived from predictive quantile-quantile (PQQ) plots (Laio and Tamea, 2007); Thyer et al., 2009; Renard et al., 2011] and has been used in numerous subsequent studies [e.g. (Mcinerney et al., 2017; Dal Molin et al., 2023; Klotz et al., 2022; Koutsoyiannis and Montanari, 2022; 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 Equation 33 of Mcinerney et al. (2017), see also (Hunter et al., 2021), defined as the standard deviation of the predictive distribution averaged over the time steps in the evaluation period, and scaled by the average observed flow. Better precision corresponds to lower values of this metric.

Volumetric bias is included to examine the long-term water balance behavior of the predictions. The volumetric bias metric is also taken from Equation 34 of Mcinerney et al. (2017), defined as the error in the total volume of the predictive distribution (averaged over all replicas and accumulated across all time steps and) relative to the total volume of the observed streamflow time series. Better volumetric bias corresponds to lower values of this metric.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

3.7 Performance of residuals error models across multiple metrics and catchments

A single residual error model, which represents a combination of a heteroscedastic transformation and an AR/ARMA model, might not achieve the best performance across all metrics. In addition, different residual error models may perform differently across multiple catchments. Hence, we employ a comparison procedure based on the approach of McInerney et al (2017) to identify the best residual error model for each performance metric. These residual error models are referred to as the "best-metric" models and are identified as follows.

For a given performance metric, differences between the performances of transformations in competing residual error models are checked for statistical significance:

1) identify the residual error model with the best median metric values.

- For example, if mA, mB, mC and mD are the performance metric values for corresponding residual error models A, B, C and D across the catchments, and median mA < median mB < median mC < median mD, then the residual error model A is the "best median" residual error model.
    - We apply 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
- The procedure above is carried out separately for the three performance metrics listed in (reliability, precision and volumetric bias)
    - for a given metric, we establish whether the degradation in performance incurred by a residual error model other than the best-median residual error model is statistically significant.

2) establish the set of best-metric residual error models, defined as the residual error models with performance that is statistically similar to the best median residual error models, when evaluated according to a particular performance metric.

A step by step example is given in the supplementary material.

3.8 ARMA models to represent persistence in residuals

Finally, we consider ARMA models to capture the persistence in the residual errors. Note that all ARMA models generate time series of random variables (here, residual errors) with the same marginal distribution at each time step. In this study, the marginal distribution refers to the unconditional distribution of residual errors (or, equivalently, streamflow predictions) at an individual time step, without an attempt to condition on observations at preceding time steps. These marginal distributions are by definition not affected by the persistence structure of the residuals.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

In addition, recall that the two-stage post-processor approach in Section 2.3 disregards posterior parameter uncertainty. Under this setup, it becomes possible to first compare multiple streamflow transformations (Section 3.3) using the prediction performance metrics in Sections 3.6-3.7, all of which apply to the marginal streamflow predictions at each time step. Then, in a separate second step, we compare multiple ARMA persistence models conditional on the residuals obtained from hydrological parameters estimated in Stage 1. This approach substantially simplifies our analysis by reducing the number of combinations of transformations and persistence models under consideration.

The following procedure is employed to examine persistence models at each catchment and residual error transformation:

1) Select an initial set of AR models, here AR1, AR2 and AR3;
2) Estimate the partial autocorrelation function (PACF) of the innovations. This step is implemented by back-calculating the innovations by inverting equation (6) and then estimating its PACF using the Matlab function "parcorr" (Box et al., 2015; Hamilton, 1994).
3) Compare the PACF of different AR models to establish which model generates innovations with the least amount of persistence. In other words, which AR model has innovations with PACFs closest to zero across all lags
4) Consider additional improvements (if any) of including MA persistence components; here we consider adding MA1, MA2 and MA3 components. If additional MA components do not provide major improvements in the PACF of the innovations, we conclude that the AR model is effectively capturing the persistence

Albeit subjective, this comparison procedure enables us to identify a parsimonious model for capturing persistence in residual errors.

4 Results

4.1 Comparison of residual error models for all flows

Table 1 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 1 compares the reliability metrics of residual error models. 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 ($\lambda = 0$, SLS) yields the worst reliability. These results are similar to the earlier work of McInerney et al (2017) with daily data which also favored Log, BC02 and BC05 transformations.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

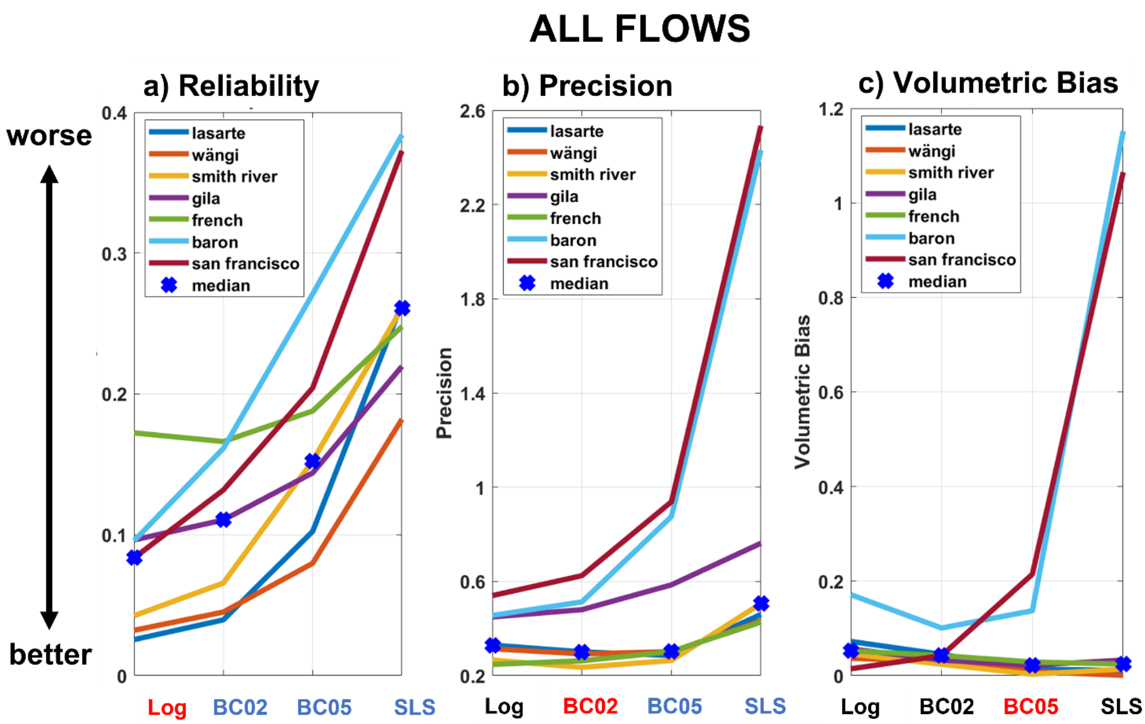


Figure 1. Performance of residual error transformations across all catchments for all flows (i.e., no stratification). Log transformation provides the best reliability. The best median transformation is indicated in red. Blue x-label indicates transformations that perform statistically significantly worse (95% confidence) than the best median transformation in terms of paired Wilcoxon test of the metric values. Bold black means there is not statistically significant difference between the transformation and the best median transformation.

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 the Log, BC02, BC05 and SLS transformations. The S shape of the PQQ plots for BC05 and SLS indicates an overestimated uncertainty when all flows are considered. This behavior 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 behavior is similar to McInerney et al [2017].

Figure 2 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.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

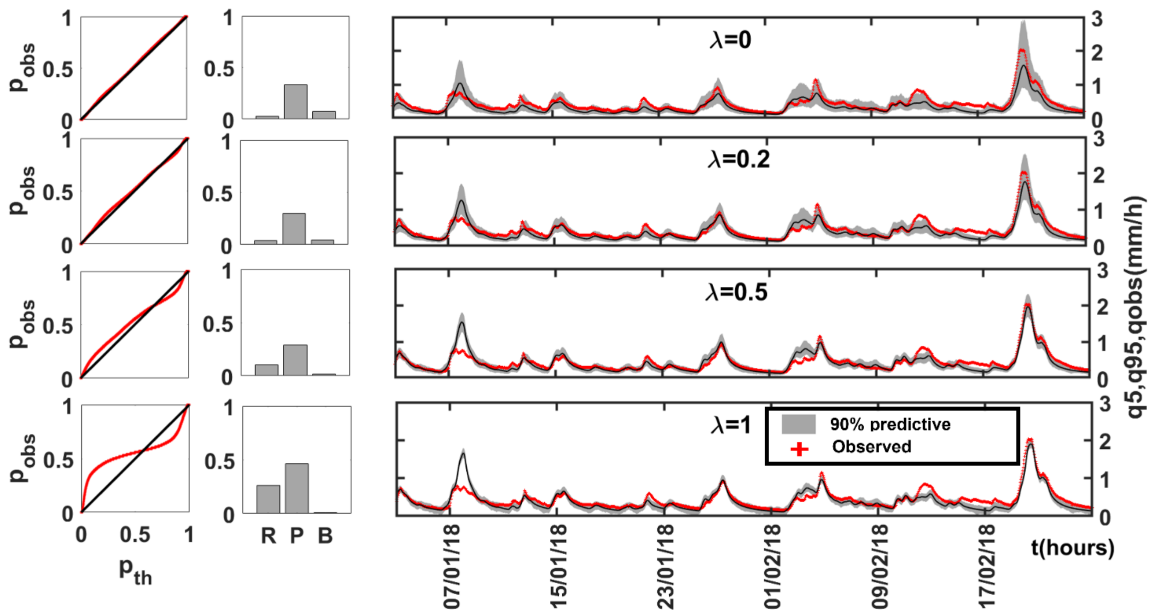


Figure 2. Illustration of probabilistic predictions in the Lasarte catchment. The results for residual error models are arranged row-by-row, for λ from 0 to 1. The left column shows the PQQ plots (Section 3.6.1). The bar plots in the middle column show the values of Reliability, Precision, and Volumetric 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 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, whereas $\lambda = 0$-0.2 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. 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 the SLS transformation has tighter predictive limits for high flows (i.e. lower uncertainty) and wider predictive limits for low flows (i.e. higher uncertainty). Further, the bar plots show that precision is worst for SLS.

Table 2. Summary of Best-Median residual error model and Best-Metric residual error model across all catchments

Metric	Best metric residual error model (best median residual error model in red)
Reliability	Log

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Precision	**BC02**, Log
Volumetric Bias	Log, BC02, **BC05**, SLS

4.1.3 Volumetric Bias

Figure 1c compares the volumetric bias metrics of residual error models. Figure 1c and Table 2 show that BC05 is the best median model (similar to McInerney et al 2017) and the best metric models are Log, BC02 and SLS. But Log, BC02 and SLS
390 are not statistically significantly worse than BC05.

The bar plots in Fig. 2 show that volumetric bias is worse for SLS than for Log, but as shown in Table 2 the difference is not statistically significant.

4.2 Comparison of residual error models for top 5% streamflow

Figure 3a shows that BC05 has slightly better reliability than the other transformations and SLS has slightly better precision
395 and bias. The median precision (Fig. 3b) and volumetric bias (Fig. 3c) improve very slightly from Log to SLS. However, these differences are found to be not statistically significant, with considerable noise across the catchments.

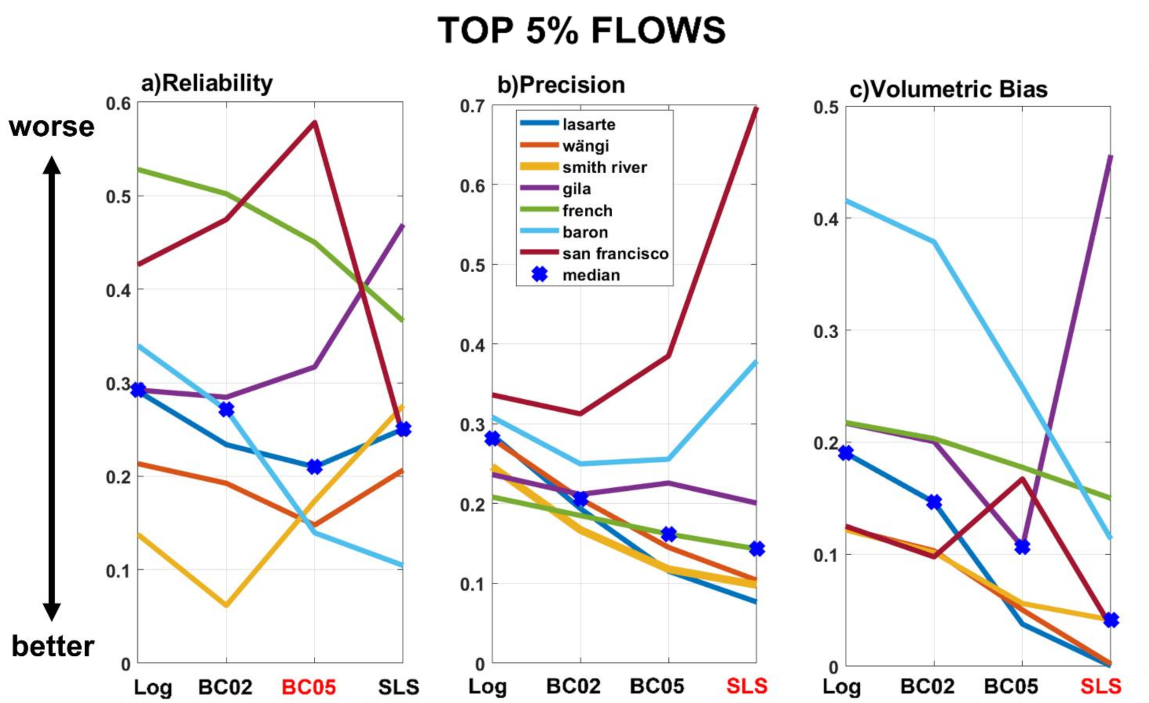


Figure 3. Performance of residual error models across all catchments for high flows (top 5% of the streamflow). Reliability is comparable for the transformations. Precision and volumetric bias improve monotonically from Log to SLS, but these changes are

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

minor and are not statistically significant. The best median transformation is indicated in red. The transformations that are not statistically significantly better than the best median transformation are highlighted in bold.Persistence model: PACF analysis of innovations

Figures 4 and 5 illustrate the comparison of AR1, AR2 and AR3 models using PACFs for Smith River and Lasarte. Persistence is better captured by the AR2 model than the AR1 model. In some catchments AR3 captures the persistence better than AR2. For lags higher than 3, the PACF is generally stable around the 95% confident bands.

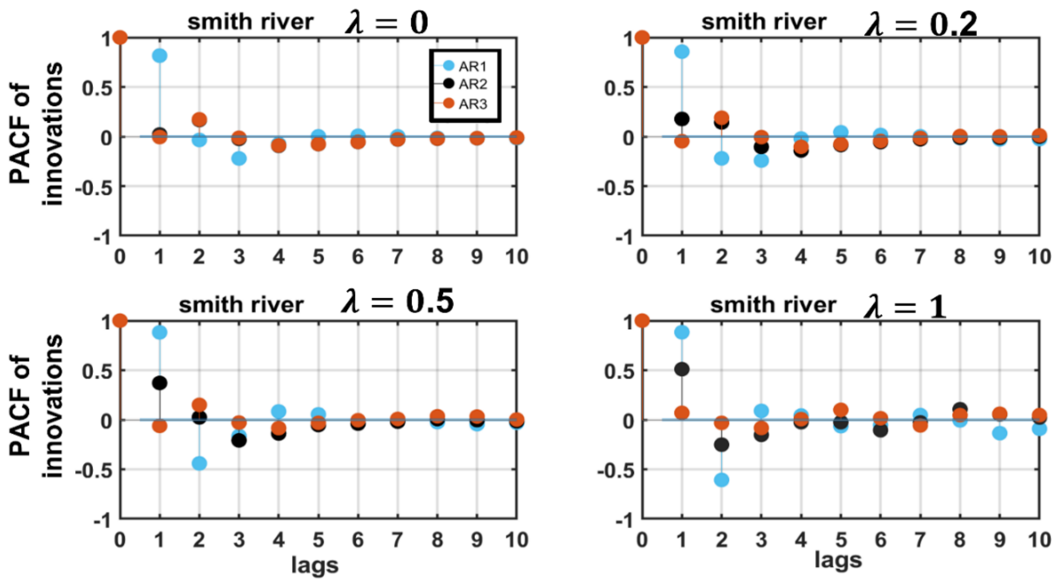


Figure 4. Comparison of AR models - Illustration 1. Partial autocorrelation of innovations when using the AR1, AR2 and AR3 persistence models in the Smith River catchment. It can be seen that the AR3 model provides the greatest reduction in the persistence of the innovations, for all values of λ.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

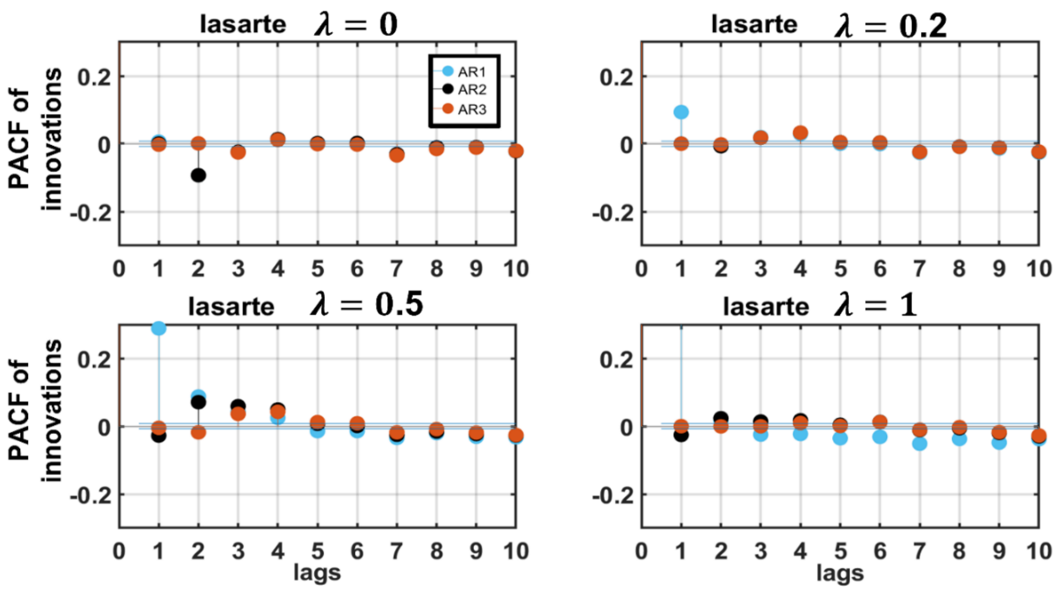
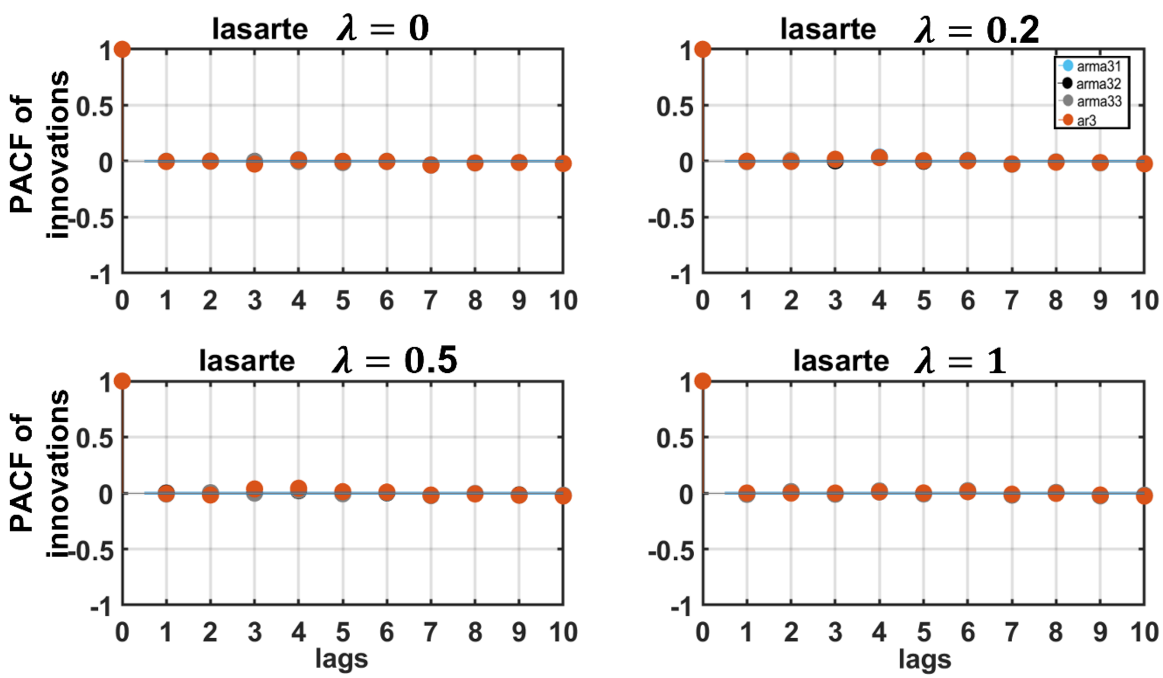


Figure 5. Comparison of AR models - Illustration 2. Lasarte catchment. Here we see a different behavior than in the Smith River catchment, with AR1-AR3 models all providing comparable reduction in PACF except when λ=0.5

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



https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Figure 6. Representative behavior resulting from including MA terms in the persistence model, illustrated using the PACF of innovations in the Lasarte catchment. It can be seen that MA terms provide little additional benefit in removing persistence beyond the AR model.

4.3 Performance at aggregated scales

The probabilistic predictions are now evaluated under temporal aggregation from hourly to daily and monthly scales. Table 3 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 underestimated uncertainty.

For example, for the log transformation, the average reliability and precision barely change from hourly to daily: reliability is 0.078 at hourly time scale and 0.085 at daily; and precision is 0.374 and 0.35, respectively. However, further aggregation to monthly time scale degrades reliability more substantially, to 0.2.

These trends hold largely 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 0.126 and precision is 0.359; while for the SLS, reliability is 0.308 and precision is 0.787.

5 Discussion

5.1 Residual error model performance for all flows

When considering performance across all flows, our results show that the transformations that provide the best median residual error models are as follows: Log transformation is the best for reliability, BC02 is best for precision and BC05 is best for volumetric bias. McInerney et al. [2017] found similar results at the daily scale.

In addition, similar to the study by McInerney et al. [2017], 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. We see that reliability improves monotonically as λ is reduced, with the best reliability achieved by the Log transformation. The consistency of findings at the hourly and daily scales is re-assuring and can be expected to facilitate the selection of residual error models for practical work.

In practice, a precise prediction is desirable only if it is also is reliable [Klotz et al. 2022]. A precise but unreliable prediction is overconfident and therefore the spread of the prediction bounds will be misleading, e.g. over-confident [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 Figure 1 versus Figure 2 in McInerney et al. [2017].

For volumetric bias, our Figure 1 is slightly different from Figure 2 in McInerney et al. [2017]. McInerney et al. [2017] show that BC0 to BC05 are better, whereas in our study none of the transformations are statistically significantly worse than the others. SLS has the best median performance and in McInerney et al. [2017], SLS has the worst median performance between

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Log, BC02, BC05 and SLS. 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 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]. Some other models in the published literature, e.g., the Countable Mixtures of Asymmetric Laplacians, CMAL method, (Klotz et al., 2022), also 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, but we have not done so based on the experience reported in McInerney et al. [2017] – the parameter inference becomes dominated by low flows which pushes λ into negative values, which 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 Residual error model performance for top 5% flows

For the top 5% of the flows, SLS has the best median performance for precision and volumetric bias, and BC05 has the best median performance for reliability. However, none of the transformations is statistically significantly worse than the others. This result is important because it suggests that residual error models based on transformations recommended across the full range of flows are not significantly disadvantaged when applied to high flows, thus avoiding the need to change to SLS models when high flows are of interest.

Note that "high" flows could be classified using different criteria than our 5% threshold. Our purpose here was to explore the performance of streamflow transformations under a commonly used high flow criterion. 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, including variance-based approaches 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 persistence. Hourly to daily aggregation has little effect on performance (as reflected in the reliability and precision metrics – see Table 3), which 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 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.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

Table 3. Mean reliability and precision across catchments for hourly daily and monthly scales, for the Log Transformation

	Log							
	reliability				precision			
	hourly	daily	monthly	Average(h,d,m)	hourly	daily	monthly	Average (h,d,m)
AR1	0.078	0.086	0.169	0.111	0.378	0.360	0.242	0.327
AR2	0.078	0.085	0.224	0.129	0.372	0.343	0.162	0.292
AR3	0.078	0.083	0.216	0.126	0.372	0.346	0.169	0.359
Average AR1, AR2, AR3	0.078	0.085	0.203		0.374	0.350	0.202	

	BC02							
	reliability				precision			
	hourly	daily	monthly	Average(h,d,m)	hourly	daily	monthly	Average (h,d,m)
AR1	0.102	0.104	0.154	0.120	0.390	0.371	0.239	0.333
AR2	0.103	0.099	0.226	0.143	0.387	0.352	0.146	0.295
AR3	0.103	0.100	0.211	0.138	0.387	0.355	0.157	0.300
Average AR1, AR2, AR3	0.103	0.101	0.197		0.388	0.359	0.181	

	BC05							
	reliability				precision			
	hourly	daily	monthly	Average(h,d,m)	hourly	daily	monthly	Average (h,d,m)
AR1	0.158	0.149	0.189	0.165	0.503	0.475	0.274	0.417
AR2	0.163	0.142	0.262	0.189	0.508	0.441	0.146	0.365
AR3	0.163	0.146	0.233	0.181	0.507	0.449	0.166	0.374
Average AR1, AR2, AR3	0.161	0.146	0.228		0.506	0.455	0.195	

	SLS							
	reliability				precision			
	hourly	daily	monthly	Average(h,d,m)	hourly	daily	monthly	Average (h,d,m)
AR1	0.272	0.260	0.364	0.299	1.063	0.956	0.546	0.855
AR2	0.276	0.264	0.382	0.307	1.079	0.790	0.394	0.754
AR3	0.276	0.262	0.387	0.308	1.079	0.848	0.433	0.787
Average AR1, AR2, AR3	0.272	0.260	0.364		1.063	0.956	0.546	

The finding that hourly to daily aggregation largely preserves performance also suggests that the persistence of the residual errors is represented sufficiently well across these two time scales. It was shown in previous work that poor representation of

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

persistence results in loss of reliability of aggregated probabilistic predictions (Evin et al., 2014). For example, if persistence is ignored altogether, the uncorrelated errors in the predictions cancel out resulting in grossly under-estimated uncertainty. For this reason, seamless prediction methods in the literature have paid considerable attention to the representation of persistence – e.g. McInerney et al. 2020 (MUTHRE paper) employed a high-order AR model with longer term memory 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 persistence – the changes in processes affecting aggregation from daily to monthly scales are more substantial than changes affecting aggregation from hourly to daily scales. This question could be explored using the longer-term persistence terms included in the MuTHRE model of (Mcinerney et al., 2022).

5.4 Practical recommendations for selecting streamflow transformations at hourly scale

In this work we have focused on simple residual error models and performance evaluation following the earlier work of [McInerney et al. 2017] but applied at the hourly time scale.

In practical applications, the selection of a transformation for 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 suggest the following recommendations for hourly predictions:

a) If reliability is the highest priority: we recommend the Log transformation, at the expense of worse precision and volumetric bias than the BC05.
b) If precision is the highest priority: still recommend the Log transformation because it is not statistically significantly worse than the BC02 transformation. Note that this choice comes at the expense of worse volumetric bias than the BC05 transformation. This recommendation is similar to McInerney et al. 2017, who found that the BC02 transformation was the best for precision, while its other metrics were statistically significantly worse.
c) If low volumetric bias is the highest priority: recommend the Log or BC02 transformations as they are not significantly worse than the BC05 and SLS transformations. The BC05 transformation achieves the best median volumetric bias, but the Log, BC02 and SLS transformations are not significantly worse. Alternatively, the BC02 transformation offers the second best reliability and the best precision. In other words, while the BC05 transformation achieves the lowest bias, this benefit comes at the expense of reduced reliability and precision. Note that McInerney et al. (2017) also found that the BC05 transformation was the best-performing metric transformation for volumetric bias but, unlike in this study, the other transformations performed significantly worse.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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 subtle 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 reflecting this behavior 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 Amman 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 considered 7 catchments, which is relatively fewer catchments than in similar studies at the daily scale, e.g. (Li et al., 2016; Hunter et al., 2021; 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 hydroclimatologies, 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 employed an evaluation approach based on statistical performance metrics (namely reliability, precision and volumetric bias), applied separately to the entire hydrograph and then to the top 5% of the 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., 2017, McInerney et al. 2019) 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 terms 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.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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 Box-Cox-type and ARMA methods considered in this work.

6 Conclusions

This study explores hourly scale streamflow predictions and associated uncertainty estimation, using comparatively simple residual error modelling approaches.

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, while its precision and volumetric bias are not statistically significantly worse than for the Box Cox transformations with power parameter of 0.2 and 0.5 respectively. These findings are 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.

When stratifying performance evaluation to the top 5% flows, the medians of precision and volumetric bias over the case study catchments improve monotonically from Log to SLS. However, these improvements are very minor and not statistically significant, well within the substantial variability across catchments.

The treatment of persistence in the errors is also an important consideration, and can be implemented using autoregressive models. Comparing AR1, AR2 and AR3 models using PACFs, the persistence is much better captured by the AR2 model than by the AR1 model. In some catchments AR3 captures the persistence better than AR2. 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. We find that, 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 results in 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 and performance metrics.

7 Data availability

The study data are deposited in https://doi.org/10.5281/zenodo.18379721.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

8 Author contribution

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.

9 Competing interests

At least one of the (co-)authors is a member of the editorial board of Hydrology and Earth System Sciences

10 Acknowledgements

The lead author acknowledges support as an academic guest at ETH Zurich and as a research visitor at Eawag, where part of this work was carried out. The lead author also acknowledges partial financial support from the Government of Cantabria through the FÉNIX programme. The authors thank Andreas Scheidegger and Grey Nearing for insightful discussions on uncertainty at the hourly time scale.

11 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, 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, Hydrology and Earth System Sciences, 23, 2147-2172, 10.5194/hess-23-2147-2019, 2019.

Bartens, A., Shehu, B., and Haberlandt, U.: Flood frequency analysis using mean daily flows vs. instantaneous peak flows, Hydrol. Earth Syst. Sci., 28, 1687-1709, 10.5194/hess-28-1687-2024, 2024.

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, 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, 108-111, 10.1038/s41586-019-1495-6, 2019.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M.: Time series analysis: forecasting and control, John Wiley & Sons2015.

Box, G. E. P. and Cox, D. R.: An Analysis of Transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211-243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x, 1964.

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 Resources Research, 44, 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 Resources Research, 48, https://doi.org/10.1029/2011WR011721, 2012.

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

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, Hydrology and Earth System Sciences, 24, 1319-1345, 10.5194/hess-24-1319-2020, 2020.

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, Scientific Data, 11, 879, 10.1038/s41597-024-03706-1, 2024.

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 Resources Research, 50, 2350-2375, 10.1002/2013wr014185, 2014.

Fill Heinz, D. and Steiner Alexandre, A.: Estimating Instantaneous Peak Flow from Mean Daily Flow Data, Journal of Hydrologic Engineering, 8, 365-369, 10.1061/(ASCE)1084-0699(2003)8:6(365), 2003.

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.

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, Engineering Geology, 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, Hydrology and Earth System Sciences, 25, 2045-2062, 10.5194/hess-25-2045-2021, 2021.

Hamilton, J. D.: Time Series Analysis, Princeton University Press, 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, Journal of Hydrology, 603, 10.1016/j.jhydrol.2021.126578, 2021.

Kirchner, J. W.: A double paradox in catchment hydrology and geochemistry, Hydrological Processes, 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, Hydrological Processes, 18, 1353-1359, 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, 10.5194/hess-26-1673-2022, 2022.

Koutsoyiannis, D. and Montanari, A.: Bluecat: A Local Uncertainty Estimator for Deterministic Simulations and Predictions, Water Resources Research, 58, 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, 10.5194/hess-11-1267-2007, 2007.

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, Hydrology and Earth System Sciences, 20, 3561-3579, 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, Journal of Hydrology, 555, 586-599, 10.1016/j.jhydrol.2017.10.057, 2017.

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

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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 Resources Research, 55, 11035-11060, 10.1029/2018wr024148, 2019.

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 Resources Research, 53, 2199-2239, 10.1002/2016wr019168, 2017.

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 Resources Research, 56, 10.1029/2019wr026979, 2020.

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

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, 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, Journal of Hydrology, 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, Journal of Hydrology, 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, Environmental Modelling & Software, 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, 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, Geophysical Research Letters, 44, 10,387-310,395, https://doi.org/10.1002/2017GL074882, 2017.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resources Research, 42, https://doi.org/10.1029/2005WR004820, 2006.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 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, 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 Resources Research, 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, 10.5194/nhess-24-3381-2024, 2024.

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 Resources Research, 46, 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 Resources Research, 47, 10.1029/2011wr010643, 2011.

Saravanapavan, T., Yamaji, E., Voorhees, M., and Zhang, G.: Using Hydrology as a Surrogate in TMDL Development for Impairments Caused by Multiple Stressors, 2014.

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 Resources Research, 46, 10.1029/2009wr008933, 2010.

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, 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 Resources Research, 60, e2023WR036710, https://doi.org/10.1029/2023WR036710, 2024.

https://doi.org/10.5194/egusphere-2026-483
Preprint. Discussion started: 4 February 2026

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, Journal of Hydrology, 575, 1031-1040, 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, 10.5194/hess-19-3951-2015, 2015.

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 Resources Research, 44, https://doi.org/10.1029/2007WR006716, 2008.

Yin, J., Gao, Y., Chen, R., Yu, D., Wilby, R. L., Wright, N., Ge, Y., Bricker, J. D., Gong, H., & Guan, M. (2023). Flash floods: Why are more of them devastating the world's driest regions? Nature, 615(7951), 212–215. *doi: https://doi.org/10.1038/d41586-023-00626-9*

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 Resources Research, 59, e2022WR033226, https://doi.org/10.1029/2022WR033226, 2023.