HYDROLOGICAL SCIENCES JOURNAL
2019, VOL. 64, NO. 4, 455–472
https://doi.org/10.1080/02626667.2019.1585858

# Systematic increase in model complexity helps to identify dominant streamflow mechanisms in two small forested basins

Paula C. David [a], Debora Y. Oliveira [a], Fernando Grison [b], Masato Kobiyama [c] and Pedro L. B. Chaffe [d]

[a]Graduate Programme in Environmental Engineering, Federal University of Santa Catarina, Florianópolis, Brazil; [b]Academic Coordination, Federal University of Fronteira Sul, Chapecó, Brazil; [c]Hydraulic Research Institute, Federal University of Rio Grande do Sul, Porto Alegre, Brazil; [d]Department of Sanitary and Environmental Engineering, Federal University of Santa Catarina, Florianópolis, Brazil

## ABSTRACT

This study shows how the use of increasing model complexity allows us to hypothesize about dominant streamflow mechanisms in two small Brazilian forested basins. Nine different structures from SUPERFLEX, an objective framework to systematically increase hydrological model complexity, were tested and we extended the flexible modelling methodology to error models as well. We show that applying a rigorous methodology in a model evaluation framework, with residual analysis and control of model complexity, is essential for testing a model as a hypothesis for dominant hydrological controls. Our results indicate that the model architecture was more important than the increase in the number of model parameters. Better performing models were those with a parallel structure, which confirms our *a priori* belief about the dominant runoff mechanisms of the studied catchments, characterized by a rapid response to rainfall, but also a constant river discharge fed by water storage on the thick soil layer.

**ARTICLE HISTORY**
Received 31 July 2018
Accepted 19 January 2019

**EDITOR**
R. Woods

**ASSOCIATE EDITOR**
S. Lyon

**KEYWORDS**
SUPERFLEX; model complexity; DREAM; Bayesian inference

## Introduction

The connectivity between basin compartments and their dependencies on storage thresholds, soil properties and topography significantly influence the hydrological behaviour at the basin scale (McGuire and McDonnell 2010). Our insufficient knowledge of those essential aspects of the system (e.g. internal organization and the ecosystem's ability to manipulate the system in response to temporal dynamics) corresponds to a significant part of the uncertainty in hydrological models (Savenije and Hrachowitz 2017). Therefore, those interactions should be considered in the conceptualization of hydrological models and its structures (Fenicia *et al.* 2011).

Conceptual hydrological models usually have a fixed structure which may be an important source of uncertainty (Butts *et al.* 2004, Poncelet *et al.* 2017). Different basins with distinct hydrological dynamics are better represented by different structures of conceptual models, which indicates a connection between basin-scale properties and the appropriate use of model structures (Van Esse *et al.* 2013, Fenicia *et al.* 2014). In order to deal with the limitation of fixed model structures, the use of flexible model structures has been recommended. For example, Clark *et al.* (2008) proposed the Framework for Understanding Structural Errors (FUSE), which combines four existing hydrological models into several model structures, and suggested that the choice of structure is as important as that of model parameters. Fenicia *et al.* (2011) and Kavetski and Fenicia (2011) proposed a flexible modelling structure called SUPERFLEX, which is based on generic blocks, such as reservoirs, joints and propagation functions, that can be assembled in different ways.

Increasing complexity in the model structure (Van Esse *et al.* 2013, Fenicia *et al.* 2014, Orth *et al.* 2015, Boer-Euser *et al.* 2017) or in the spatial scale (Van Der Linden and Woo 2003, Li *et al.* 2015) does not guarantee improved simulations. It is not the number of parameters that determines the capacity of the model to reproduce the responses of a basin, but rather the role of those parameters, the processes they represent, and their impacts on the basin response (Fenicia *et al.* 2008). A higher number of parameters can lead to over-fitting (Orth *et al.* 2015), since the additional parameters may be fitted to noise of the observed data and thus lead to poor predictive performance (Schöniger *et al.* 2014, Lever *et al.* 2016). Overfitted models might perform better in the calibration dataset, but they are likely to perform poorly in the validation (Perrin *et al.* 2001, Lever *et al.* 2016). Model complexity

**CONTACT** Pedro L. B. Chaffe ✉ pedro.chaffe@ufsc.br
© 2019 IAHS

control reduces parameter equifinality and helps to identify robust models, allowing hydrological generalization and classification (Schoups *et al.* 2008). Also, it avoids over-fitting and low parameter sensitivity (Schöniger *et al.* 2014). As a result, the best model will be the one with a better balance between goodness of fit and complexity.

Violations of the assumptions about model residuals are known to lead to biased parameter estimates (Schoups and Vrugt 2010), and thus have a high potential to misguide model selection. For that reason, it is important that these assumptions are verified *a posteriori* (e.g. Thyer *et al.* 2009, Schoups and Vrugt 2010, Smith *et al.* 2010, 2015, Kavetski *et al.* 2011). Despite those previous findings, the adoption of a formal model residual treatment is still incipient in the hydrological modelling literature.

In this study we used a flexible hydrological modelling strategy to identify model structures with a better correspondence with catchment behaviour and hence hypothesize about dominant runoff generation mechanisms. In order to prevent biased results in the choice of both model parameters and model structures, we extended the flexible modelling methodology to error models as well and combined it with uncertainty analysis and control of model complexity. As a case study, the proposed methodology was applied to the rainfall-runoff modelling of two forested basins located in the southern region of Brazil. We show that, even for a relatively small dataset, the adoption of this rigorous methodology is useful for model identification.

## Materials and methods

### Study site

Two small experimental catchments were used in this work: the Bugres River basin (11.45 km$^2$) and the Saci River basin (0.102 km$^2$) (Fig. 1). The study of experimental catchments is important because the observed behaviour can be linked to specific catchment characteristics. The catchments are located in the northern region of the state of Santa Catarina (Southern Brazil). Both have precipitation and streamflow data with high temporal resolution (10 min). Due to the costs of maintenance and limited funding cycle, only relatively short periods of data were available. The precipitation and streamflow data at the Bugres River basin were collected by Grison *et al.* (2014) from 11 May 2011 to 1 July 2014. The rating curve has an upper level limit, which corresponds to the bankfull of the monitoring section. Discharge values related to water levels larger than this bankfull level have low reliability and their use increases the uncertainty of the data. Therefore, points with levels above these limits were excluded, resulting in two periods of continuous records (without excluded data): from 4 April to 9 July 2012 and from 20 September 2012 to 6 February 2013. The basin elevation varies from 820 to 981 m a.s.l. and the elevation of the meteorological station is 790 m a.s.l. (Grison *et al.* 2014). We used streamflow data from the Saci River basin from 3 October to 17 November 2008 (Chaffe *et al.* 2010). The mean elevation of this basin is 960 m a.s.l. (Santos 2009). Precipitation was measured at a meteorological station located 1 km away from the basin outlet, with an elevation of 869 m a.s.l. (Chaffe *et al.* 2010). The elevation variation is not large; therefore we considered that the precipitation is constant over the basins.

Daily potential evapotranspiration was calculated using the modified Penman method (Doorenbos and Pruitt 1977), which requires daily data of temperature, incident radiation, relative humidity, and average wind speed at 2 m above the ground surface. The meteorological data were obtained at the Feio weather station, which is located 1 km south of the Saci River basin. Daily potential evapotranspiration data were transformed to 10 min – the temporal resolution of the model input data – considering that evapotranspiration behaves as a sinusoidal function from 06:00 to 18:00 h, corresponding to 90% of the total potential evapotranspiration of the day, and has constant values during the rest of the day. This type of transformation of daily potential evaporation data into sub-daily data using a sinusoid function was also employed in other works (e.g. Fenicia *et al.* 2006, Vaché and McDonnell 2006).

### Hydrological model framework

The SUPERFLEX framework (Fenicia *et al.* 2011, Kavetski and Fenicia 2011) consists of reservoirs and connections that conceptualize the storage and release of water. They can represent elements such as interception, surface flow, soil moisture and groundwater, among others, presenting a linear or nonlinear outflow. The models can also be classified according to the different connectivity hypotheses of the flow paths (i.e. serial or parallel structures), which is quite similar to the Tank Model proposed by Sugawara (1961, 1995)).

The SUPERFLEX framework has been tested and found to be suitable for several catchments with areas ranging from 0.04 to 10 009 km$^2$ (Kavetski and Fenicia 2011, Van Esse *et al.* 2013, Fenicia *et al.* 2014, Gao *et al.* 2014). Fenicia *et al.* (2014) found with the SUPERFLEX framework that experimental basins with a “vertical”

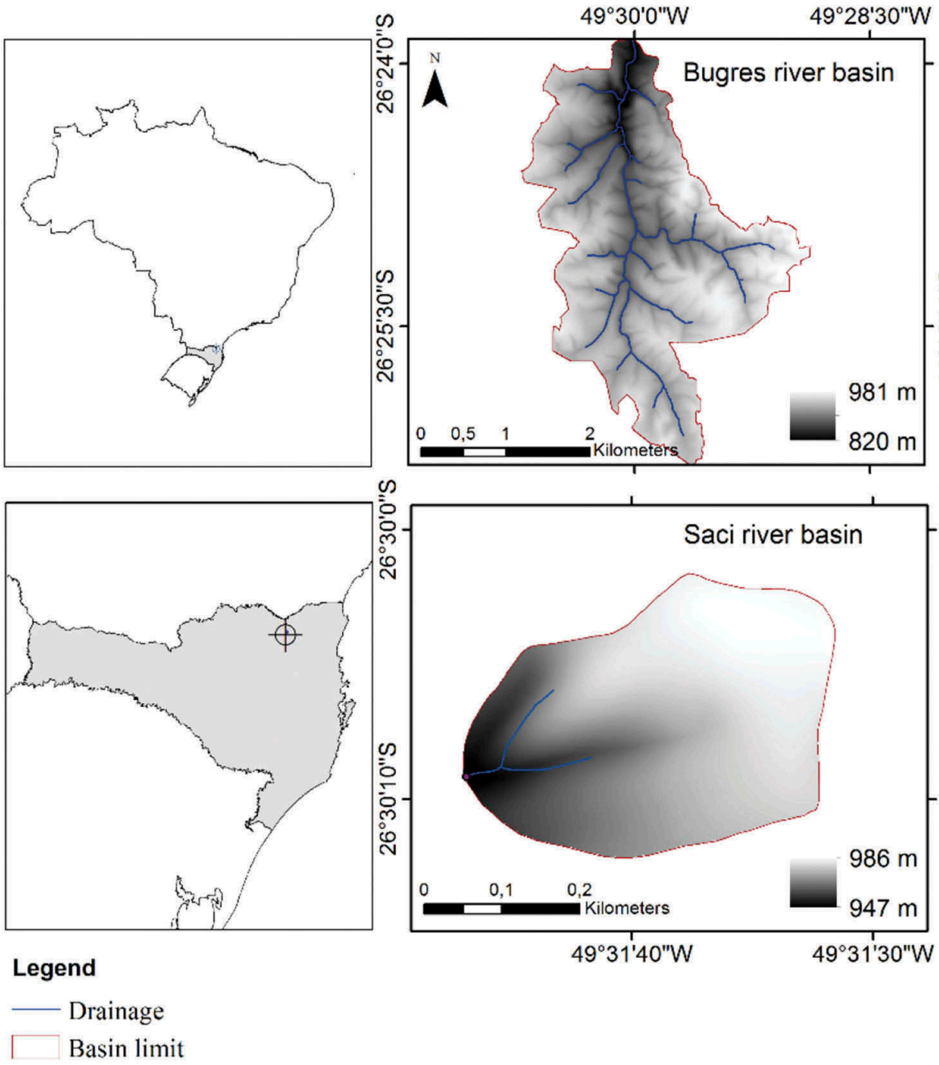


Figure 1. Location of the Saci and Bugres river basins.

behaviour (water flow) are best represented by models with parallel connections. Those connections in parallel consider the distribution of precipitation in fast and slow reservoirs. However, “horizontal” basins were best represented by models with serial connections. Van Esse *et al.* (2013) used the SUPERFLEX in 237 French basins and found that the inclusion of a slow reservoir representing the subsurface flow improves the model performance in basins with dominant groundwater, since it allows for the independence of the fast and slow flows.

We used nine different structures of the SUPERFLEX framework proposed by Fenicia *et al.* (2011) and Kavetski and Fenicia (2011) (Fig. 2): eight were the same as in Fenicia *et al.* (2014) – M03, M04, M06, M07, M08, M09, M11 and M12 – and one was a new combination of the model structures, which we called M13. Models M03, M04 and M06 have a serial structure. Model M03 has two reservoirs; the precipitation enters the unsaturated zone reservoir and the storage that exceeds a specified threshold overflows and enters the fast reservoir. Model M04 differs from M03 because the outflow of the unsaturated zone reservoir occurs according to an exponential function, rather than a threshold. The only difference between M04 and M06 is that M06 contains an interception reservoir.

Models M07, M08, M09, M11, M12 and M13 have a parallel structure. Model M07 has a riparian zone reservoir, which receives a constant fraction of the total precipitation. In M08 the precipitation is divided into fast and slow linear reservoirs. Model M09 differs from M08 by the inclusion of an unsaturated zone reservoir whose outflow is divided into fast and the slow reservoirs, which are both linear. Model M11 differs from M09 by including a nonlinear function at the outflow of the unsaturated zone reservoir. Model M12 differs from M11 by the addition of an interception reservoir.

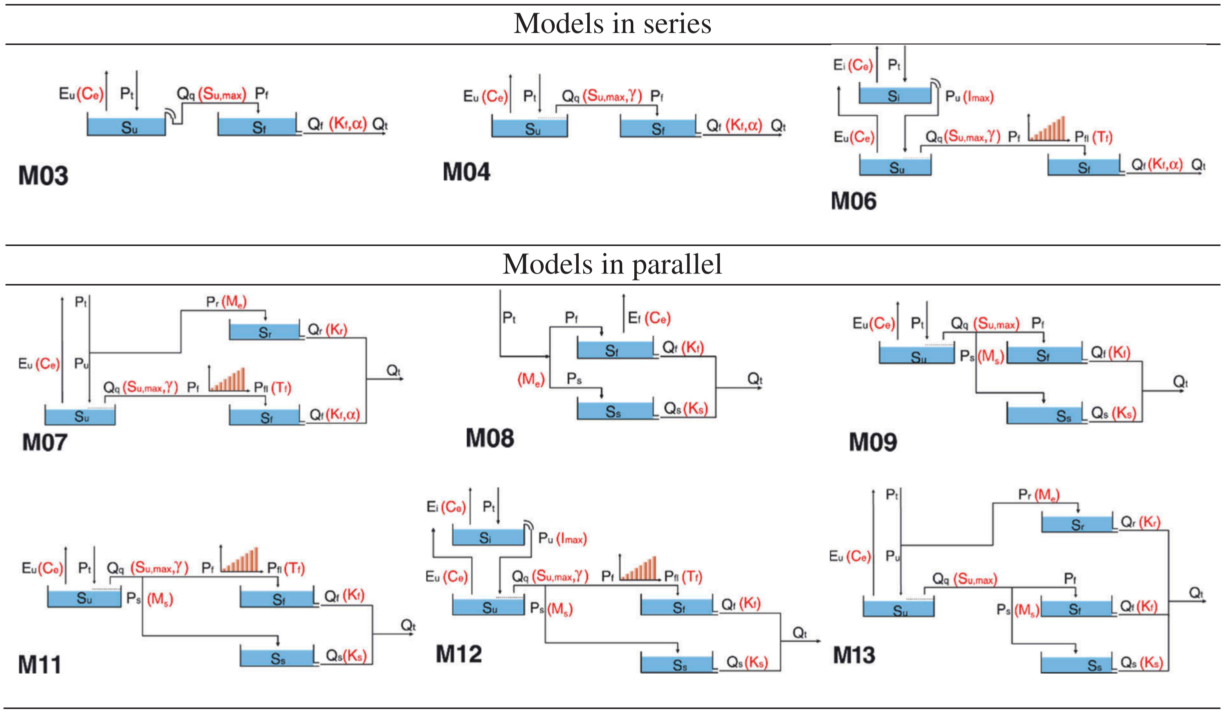


**Figure 2.** Model structures from SUPERFLEX framework considered in this study. The parameters are in red (adapted from Fenicia *et al.* 2014).

**Table 1.** Physical processes included in each model structure.

| | M03 | M04 | M06 | M07 | M08 | M09 | M11 | M12 | M13 |
|---|---|---|---|---|---|---|---|---|---|
| Unsaturated reservoir | × | × | × | × | | × | × | × | × |
| Fast reservoir | × | × | × | × | × | × | × | × | × |
| Slow reservoir | | | | | × | × | × | × | × |
| Interception reservoir | | | × | | | | | | × |
| Riparian zone reservoir | | | | × | | | | | |

Finally, model M13 is M09 with a riparian zone reservoir. Table 1 summarizes which processes (reservoirs) are considered in each structure.

The construction of the models in a controlled way allows us to attribute differences in performance to differences in model structure. One can test the influence of serial *versus* parallel connections, the importance of the interception reservoir and the linearity of the processes, for example. The models were implemented in MATLAB with a second-order accurate explicit method with adaptive time stepping (Schoups *et al.*, 2010), absolute and relative tolerances fixed at $10^{-3}$, utilizing the water balance equations and the constitutive relationships presented in Appendix A of Fenicia *et al.* (2014).

## *Calibration and uncertainty analysis*

Model calibration and uncertainty analysis were performed using the automatic calibration algorithm Differential Evolution Adaptive Metropolis (DREAM$_{(ZS)}$) proposed by Laloy and Vrugt (2012) and Vrugt (2016). DREAM uses Bayesian inference for the joint estimation of model parameter values and their uncertainty. We set up the DREAM$_{(ZS)}$ parameters so that the number of Markov chains was $N = 3$ and the number of generations $T = 15\,000$. In some cases, $T$ was increased to guarantee the convergence to a stationary distribution.

One issue of Bayesian statistics is that the construction of the likelihood function requires some assumptions about model residuals to be made *a priori*, and often these assumptions are not met or verified *a posteriori*. The violation of those premises leads to unreliable parameter and uncertainty estimates (Thyer *et al.* 2009, Schoups and Vrugt 2010, Smith *et al.* 2010, 2015, Kavetski *et al.* 2011, Oliveira *et al.* 2018). In this work, the generalized likelihood function (GL) proposed by Schoups and Vrugt (2010) was used for the inference of the hydrological model parameters. The GL relaxes the commonly assumed premises on residual errors, allowing different hypotheses about the residual model to be considered (Schoups and Vrugt 2010).

The natural logarithm of the likelihood function is used for algebraic simplicity and for having greater numerical stability:

$$\ell = n\log\frac{2\sigma_\xi\omega_\beta}{\xi+\xi^{-1}} - \sum_{t=1}^{n}\log\sigma_t - c_\beta\sum_{t=1}^{n}|a_{\xi,t}|^{\frac{2}{1+\beta}} \qquad (1)$$

**Table 2.** Assumptions of each residual model considered in this study.

| Model | Correlation | Heteroscedasticity | Distribution | Implementation |
|---|---|---|---|---|
| L1 | Independent | Homoscedastic | Gaussian | $\sigma_1 = 0$; $\beta = 0$; $\xi = 1$ |
| L2 | Independent | Heteroscedastic | Gaussian | $\beta = 0$; $\xi = 1$ |
| L3 | Independent | Heteroscedastic | SEP | $\xi = 1$ |

where $n$ is the number of discharge observations used for parameter inference; $a_{\xi,t}$, $\omega_\beta$, $c_\beta$ and $\sigma_\xi$ derive from the values of skewness $\xi$ and kurtosis $\beta$ (equations are presented in Appendix A of Schoups and Vrugt 2010).

In order to evaluate the considered assumptions, different error models with increasing complexity were tested in a systematic way with the inclusion of different parameters in the inference, as done in previous studies (e.g. Schoups and Vrugt 2010, Smith *et al.* 2015, Oliveira *et al.* 2018). The first model used (L1) has been widely used and it is the simplest one. It considers that the residuals follow a Gaussian distribution, with zero mean and constant variance, and are independent. The second error model (L2) considers that the errors are heteroscedastic. The heteroscedasticity of the residuals was considered assuming that the error standard deviation increases linearly with the simulated flow (e.g. Schoups and Vrugt 2010, Evin *et al.* 2014, Westra *et al.* 2014, Oliveira *et al.* 2018):

$$\sigma_t = \sigma_0 + \sigma_1 \hat{y}_t \quad (2)$$

where $\sigma_t$ is the standard deviation at time *t;* $\sigma_0$ is the heteroscedasticity intercept, $\sigma_1$ is the heteroscedasticity slope, and $\hat{y}_t$ is the simulated flow. The third model (L3) considers a skewed exponential power (SEP) distribution of errors. In this case, we allowed the kurtosis parameter ($\beta$) to vary, while the skewness parameter remained fixed (we only considered symmetric distributions). A summary of the error models derived from GL is presented in Table 2. Further details on the implementation of GL can be found in Schoups and Vrugt (2010). The error model parameters were inferred jointly with the hydrological model parameters. We used a uniform prior pdf for each parameter with ranges specified in Table 3.

The residuals were highly correlated when the original dataset – with discharge measurements at 10-min intervals – was used for model calibration. We tested three different approaches to handle autocorrelation: (a) consideration of an AR(1) model applied to the raw residuals, as in the original GL formulation; (b) consideration of an AR(1) model applied to standardized residuals, as suggested by Evin *et al.* (2013); and (c) a reparameterization of (b), as presented by Evin *et al.* (2014). In all three approaches, with temporal resolution of 10 min, the value of $\phi$, the parameter of the AR (1) model, converged to 1, which results in large random errors (Schoups and Vrugt 2010). We also combined the three approaches with different thinning. The results were worse for an interval smaller than 6 h, and better for intervals of 12 and 24 h. However, approach (a) may result in a poor predictive uncertainty, as suggested by Evin *et al.* (2013), and approach (b) resulted in a high correlation between $\sigma_1$ and $\phi$, as also demonstrated by Evin *et al.* (2013). Therefore, since the use of an AR(1) model alone was not enough to handle residual autocorrelation and to avoid the problems encountered when the AR(1) model was employed, we decided to consider only a thinning of the data series, as done in other hydrological studies (e.g. Westra *et al.* 2014). Other options would be possible, such as to fix the parameter of the AR(1) model to a pre-specified value (Schoups and Vrugt 2010) or to separately infer the hydrological and error model parameters (Evin *et al.* 2014).

The models were recalibrated using a dataset composed of one every $k$th discharge value. Values of $k$ equal to 6, 36 and 72 were tested (i.e. observation intervals of 1, 6 and 12 h). Error autocorrelation was

**Table 3.** Hydrological and error model parameters specifications.

| | Parameter | Description | Min | Max | Unit |
|---|---|---|---|---|---|
| Hydrological model | $C_e$ | Evaporation parameter | 0.01 | 2 | - |
| | $S_{u,\max}$ | Unsaturated reservoir storage capacity | 0.1 | 700 | mm |
| | $I_{\max}$ | Interception reservoir storage capacity | 0.1 | 500 | mm |
| | $\gamma$ | Unsaturated reservoir exponent | 0.001 | 20 | - |
| | $M_e$ | Inflow partitioning coefficient | 0.01 | 0.99 | - |
| | $M_s$ | Outflow partitioning coefficient | 0.01 | 0.99 | - |
| | $\alpha$ | Fast reservoir exponent | 1.0 | 20 | - |
| | $K_r$ | Riparian zone reservoir coefficient | 0.01 | 10 | $h^{-1}$ |
| | $K_f$ | Fast reservoir coefficient | 0.001 | 1 | $mm^{1-\alpha} h^{-1}$ |
| | $K_s$ | Slow reservoir coefficient | 0 | 1 | $h^{-1}$ |
| Error model | $\sigma_0$ | Heteroscedasticity intercept | 0 | 1 | mm $h^{-1}$ |
| | $\sigma_1$ | Heteroscedasticity slope | 0 | 1 | - |
| | $\beta$ | Kurtosis parameter | −1 | 1 | - |

significantly reduced in all the models when utilizing a thinning of 72. For the Saci River basin we obtained similar results (not shown), therefore a thinning of 72 was adopted.

The last 7500 sets of parameters sampled with the DREAM$_{(ZS)}$ algorithm were used to represent the uncertainty associated with the parameter values and to create the probabilistic streamflow simulations. The performance of each model was evaluated using three different metrics: the reliability, the precision and the volumetric bias metrics. The reliability of the probabilistic distribution was evaluated with the reliability metric (Evin *et al.* 2014, McInerney *et al.* 2017):

$$\text{Reliability}[\hat{y}, y] = \frac{2}{n}\sum_{t=1}^{n}\left|F_U\left[F_{\hat{y}(t)}(y_t)\right] - F_\Omega\left[F_{\hat{y}(t)}(y_t)\right]\right| \tag{3}$$

where $F_{\hat{y}(t)}$ is the cumulative distribution function (cdf) of the predictive distribution, $y$ is the observations, $\hat{y}$ is the simulations, $F_U$ is the cdf of the uniform distribution, and $F_\Omega$ is the empirical cdf. The precision metric is related with the width of the probabilistic predictions (McInerney *et al.* 2017):

$$\text{Precision}[\hat{y}, y] = \frac{\frac{1}{n}\sum_{t=1}^{n}\text{sdev}\hat{y}_t}{\frac{1}{n}\sum_{t=1}^{n}y_t} \tag{4}$$

where sdev$\hat{y}_t$ is the standard deviation of the probabilistic predictions at time step $t$. The volumetric bias metric evaluates the model's capacity in simulating the water balance (McInerney *et al.* 2017):

$$\text{VolBias}[\hat{y}, y] = \left|\frac{\sum_{t=1}^{n}y_t - \sum_{t=1}^{n}\hat{y}_{t,\text{mean}}}{\sum_{t=1}^{n}y_t}\right| \tag{5}$$

where $\hat{y}_{t,\text{mean}}$ is the average of the simulations at time step $t$. For all the metrics considered, the value of zero indicates the perfect performance.

### Control of model complexity

One way to control model complexity is by separating the available data in a period for calibration and a period for validation. The comparison between the models can be performed by evaluating their performance in the validation period. For the Bugres River basin, the dataset was divided into two parts – from 4 April to 9 July 2012 for calibration and from 20 September 2012 to 6 February 2013 for validation. However, for the Saci River basin, the series was not long enough to be separated into two parts. Hence, information criteria were used to control the complexity and select the best model. For the Bugres River basin performance in both validation and information criteria was used to select the best model in order to verify whether the two approaches lead to similar conclusions. As information criteria, we used the Akaike information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (Schwarz 1978). These information criteria evaluate which hypothesis (i.e. which model) is better supported by the data:

$$I_k = -2\ln L + C \tag{6}$$

where $I_k$ is the information criterion value, $L$ is the maximum likelihood of each hypothesis, and $C$ is a positive scalar that penalizes the complexity. Here note that $C = 2d$ for the AIC and $C = d\ln(n)$ for the BIC, where $d$ is the number of parameters and $n$ the number of observations. The best model among all is the one with the lowest value of $I_k$. The relative support of a model in relation to the best model (the one that has the lowest information criterion value) is calculated by the difference between the AIC value for the model considered, AIC$_i$, and the information criterion value for the model with the lowest value of AIC, AIC$_{\text{min}}$:

$$\Delta A_i = \text{AIC}_i - \text{AIC}_{\text{min}} \tag{7}$$

With this value, weights can be assigned to each of the $n_m$ models considered,

$$w_i = \frac{\exp(-\frac{1}{2}\Delta A_i)}{\sum_{j=1}^{n_m}\exp(-\frac{1}{2}\Delta A_j)} \tag{8}$$

where $w_i$ is the weight for model $i$. The weights for the BIC are calculated in the same way. The values indicate the probability of model $i$ being chosen in a different period from that used in the calibration.

## Results

### Parameter and predictive uncertainty

In order to prevent biased results in the choice of both model parameters and model structures, the comparison between the hydrological models was preceded by an analysis of the use of different error models for parameter inference. To evaluate the impact of the error model in the parameter and predictive uncertainty, we compared the posterior parameter distributions and the value of the three performance metrics (reliability, precision and volumetric bias) among them.

### Bugres River basin

For the Bugres River basin, considering the performance of the three different error models evaluated in the calibration period, L1 clearly presented poor results for precision and reliability when compared with the error models L2 and L3, and, in general, there was a slight improvement in these metrics with L3 in comparison with L2 (Fig. 3). These results are in agreement with the conclusions presented in Kavetski *et al.* (2011), Schoups and Vrugt (2010) and Thyer *et al.* (2009) that the use of a more adequate error model improves the characterization of the predictive uncertainty. However, the volumetric bias metric was overall lower (i.e. better) for L1 and L2. The analysis of the different hydrological model structures throughout the Results and Discussion sections was made with the results of the error model L3.

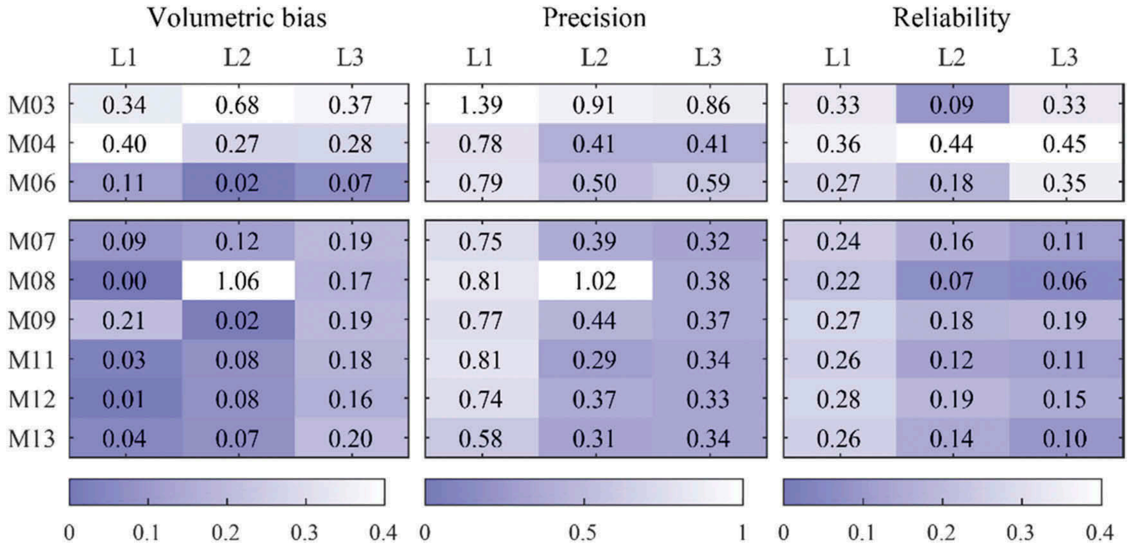
With respect to the different hydrological model structures, in general models with a parallel structure outperformed those in series considering all the metrics. Between the models with a parallel structure (M07–M13), the performance metrics values were very similar, and models M08 and M13 presented better results for reliability. Figure 4 presents the predictive uncertainty for models M04 (in series) and M07 (in parallel) obtained with the different error models.

The consideration of a threshold function for the outflow of the unsaturated zone reservoir led to poorer performance (M03 *versus* M04). Comparing M04 and M06, the inclusion of an interception reservoir improved the volumetric bias and reliability metrics. Among models M11 and M12 the inclusion of this reservoir improved the precision and volumetric bias metrics but led to poor performance considering the reliability metric. Model M07 outperformed M04 in all metrics; the addition of a riparian zone reservoir improved the representation of the discharge behaviour in this basin. The inclusion of a nonlinear outflow in the unsaturated zone reservoir (M11 versus M09) improved the performance as well.

The parameters inferred for the Bugres River basin resulted in different posterior distributions with each error model and with each hydrological model. An example of the posterior parameter distributions obtained with the three error models is presented in Figure 5 for the model M07. The $C_e$ parameter, which is an adjustment factor for the evapotranspiration, converged to its upper limit, equal to 2. The parameter $\alpha$ had values greater than 1.5 in most models, implying a nonlinearity of the surface runoff generation. The parameter $\gamma$ in most of the models was smaller than 1, indicating also a nonlinearity of the slow reservoir. These results justify the slightly worst performance of linear models, such as M09, in which these two parameters were set as 1.

### Saci River basin

For the Saci River basin, the precision and reliability metrics presented better results for most of the models when the L3 model was considered. The value of the reliability metric was up to 50% lower (better) with L3 when compared to L1. The value of the precision metric improved by up to 40%. Nevertheless, as for



| | Volumetric bias | | | Precision | | | Reliability | | |
|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L1 | L2 | L3 | L1 | L2 | L3 |
| M03 | 0.34 | 0.68 | 0.37 | 1.39 | 0.91 | 0.86 | 0.33 | 0.09 | 0.33 |
| M04 | 0.40 | 0.27 | 0.28 | 0.78 | 0.41 | 0.41 | 0.36 | 0.44 | 0.45 |
| M06 | 0.11 | 0.02 | 0.07 | 0.79 | 0.50 | 0.59 | 0.27 | 0.18 | 0.35 |
| M07 | 0.09 | 0.12 | 0.19 | 0.75 | 0.39 | 0.32 | 0.24 | 0.16 | 0.11 |
| M08 | 0.00 | 1.06 | 0.17 | 0.81 | 1.02 | 0.38 | 0.22 | 0.07 | 0.06 |
| M09 | 0.21 | 0.02 | 0.19 | 0.77 | 0.44 | 0.37 | 0.27 | 0.18 | 0.19 |
| M11 | 0.03 | 0.08 | 0.18 | 0.81 | 0.29 | 0.34 | 0.26 | 0.12 | 0.11 |
| M12 | 0.01 | 0.08 | 0.16 | 0.74 | 0.37 | 0.33 | 0.28 | 0.19 | 0.15 |
| M13 | 0.04 | 0.07 | 0.20 | 0.58 | 0.31 | 0.34 | 0.26 | 0.14 | 0.10 |

Figure 3. Performance metrics for the Bugres River basin, using error models L1, L2 and L3. Calibration performed with runoff data with thinning of 72.

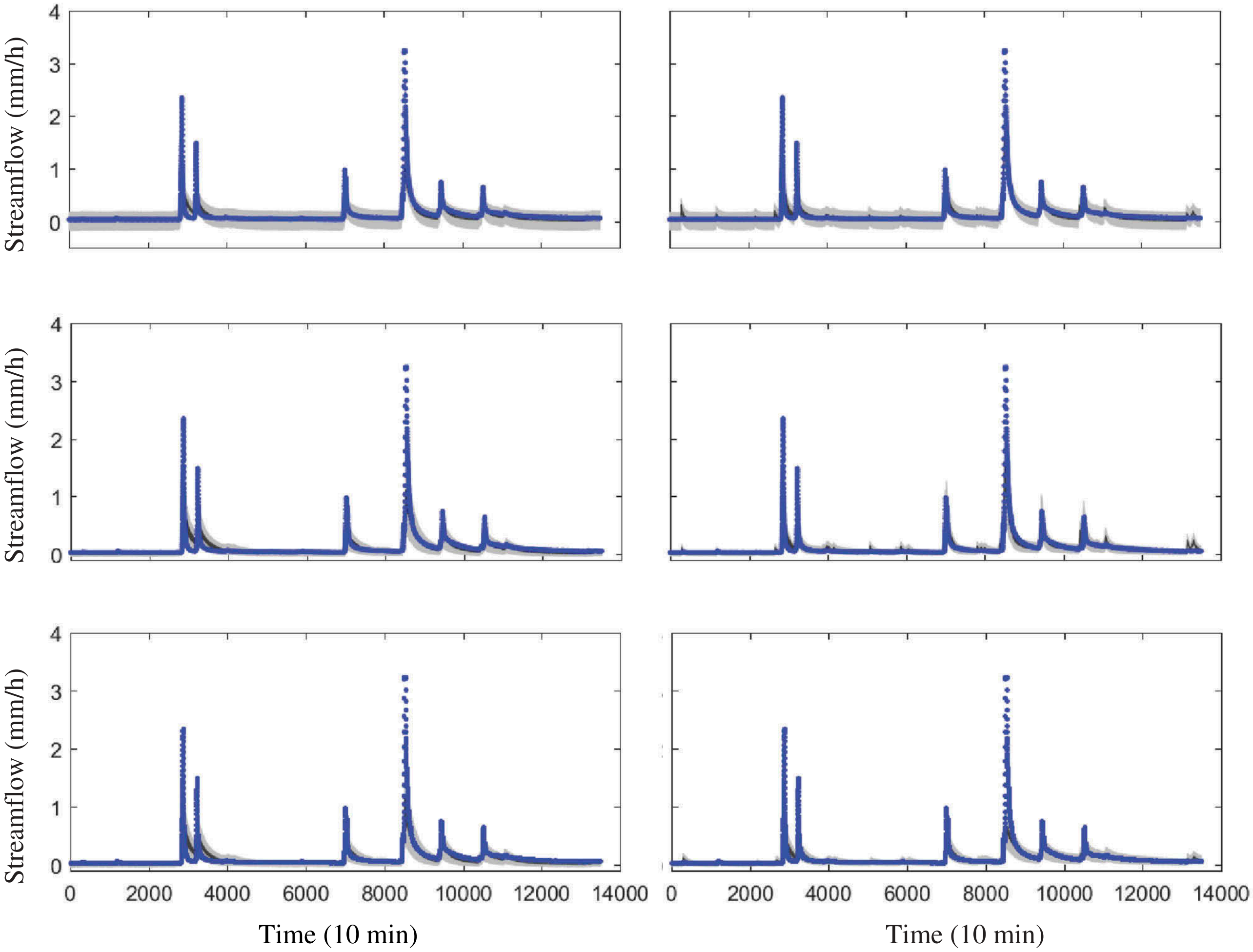


Figure 4. Observed runoff series (blue), 95% uncertainty (light grey) and uncertainty associated with the values of the parameters (dark grey) for the Bugres River basin for models M04 (left) and M07 (right), using error models L1 (top), L2 (middle) and L3 (bottom). Calibration performed with thinning of 72.

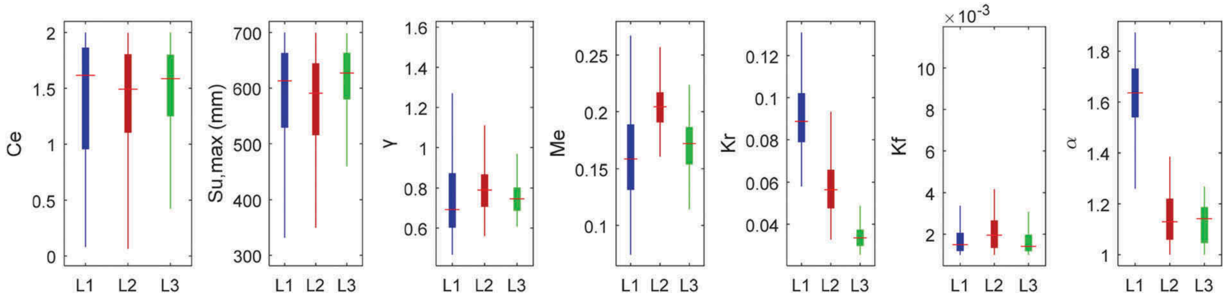


Figure 5. Distribution of the parameters inferred with each of the three error models – L1, L2 and L3 – for model M07 in the Bugres River basin. The central mark indicates the medians, the box indicates the 50% quantiles and the whiskers extend to the 95% quantiles.

the Bugres River basin, the volumetric bias metric presented better (smaller) results with L1 and L2. The comparison between the hydrological model structures throughout the Results and Discussion sections was done using the results obtained with L3.

Performance metrics for the calibration period in the Saci River basin also showed a significant difference between structures in series and in parallel, as was the case of the Bugres River basin (Fig. 6). Moreover, models with a serial structure (M03, M04 and M06)

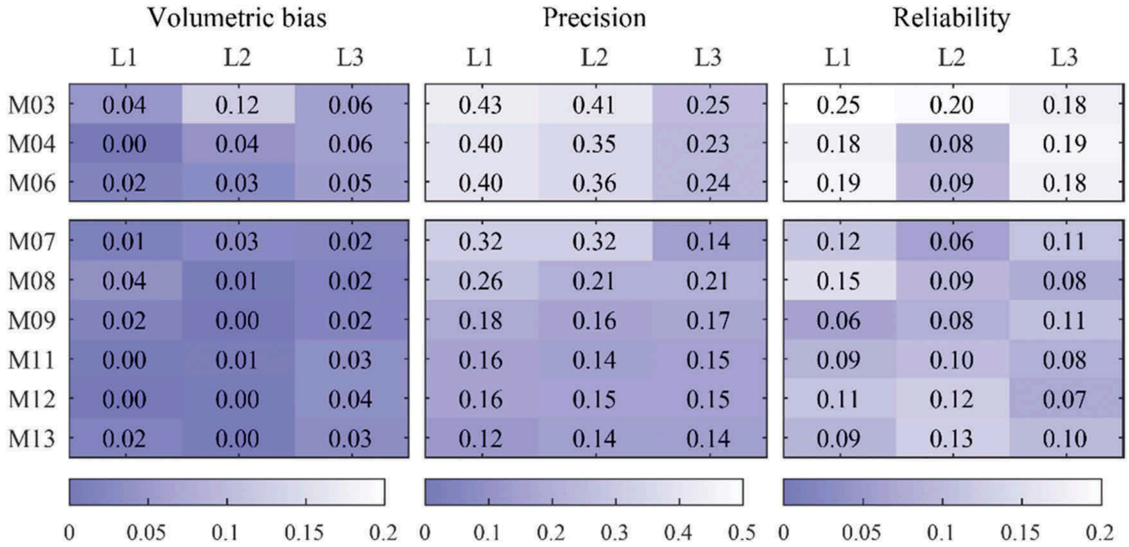


**Figure 6.** Performance metrics for the Saci River basin, using error models L1, L2 and L3. Calibration performed with runoff data of the complete series with data for 10 min and thinning of 72.

underestimated the peak flows considerably when compared with those with structure in parallel. To illustrate this result, Figure 7 presents the predictive uncertainty for models M03 (in series) and M11 (in parallel) obtained with the different error models.

Models M09, M11, M12 and M13 presented similar results for all three metrics. In the calibration the best models were those with an unsaturated zone reservoir with its outflow split between two reservoirs. The inclusion of a slow and a fast reservoir independent from each other led to better prediction of catchment outflow.

Some components from the structures did not improve the predictive uncertainty. There was some improvement in the results of M04 in relation to M03. The difference between these models is the outflow of the unsaturated reservoir: in M04 it is an exponential function and in M03 a threshold function. The inclusion of an interception reservoir (M06 *versus* M04) did not improve the results significantly, resulting in lower (better) volumetric bias and reliability values and a poorer performance in precision. M11 and M12 performed very similarly, with M12 presenting lower (better) values of reliability and precision metrics and larger (worst) volumetric bias. The addition of a riparian zone reservoir (M04 *versus* M07) increased the performance for the volumetric bias and reliability metrics. The inclusion of a nonlinear outflow in the unsaturated zone reservoir (M11 *versus* M09) improved precision and reliability but increased the volumetric bias.

As in the Bugres River basin, the parameter posterior distributions varied among both the hydrological models and the error models. As an example, Figure 8 presents the results for M11, which indicate that errors in the correct representation of the residuals or the hydrological model structure can be compensated by the distribution of the parameter values. Considering all the models, the values of $M_s$, which distributes the flow between the surface (fast reservoir) and subsurface (slow reservoir) flows, were close to 0.90; i.e. only 10% of the flow leaving the unsaturated zone reservoir goes to the slow reservoir. The $\alpha$ parameter, which represents the nonlinearity of the fast reservoir, was close to 1, which indicates that this reservoir has a linear behaviour. The $\gamma$ parameter values were larger than 4 for models M11, M12 and M13, showing a significant nonlinearity of the slow reservoir. The parameter $C_e$ converged to values close to and smaller than 1, which means that there was no need to adjust the potential evapotranspiration.

### *Analysis of model complexity*

To verify which model is better supported by the data, two complexity control methods were used: (i) separation of the data series in two parts, one for calibration and the other for validation; and (ii) two information criteria – the AIC and the BIC. The data series of the Saci River basin was not long enough to be split in two parts. Therefore, only the second method was used. For the Bugres River basin, both methods were utilized to verify whether the two approaches led to similar conclusions, which would indicate whether the use of information criteria alone might be enough to identify the best model. Hence, it would support the model

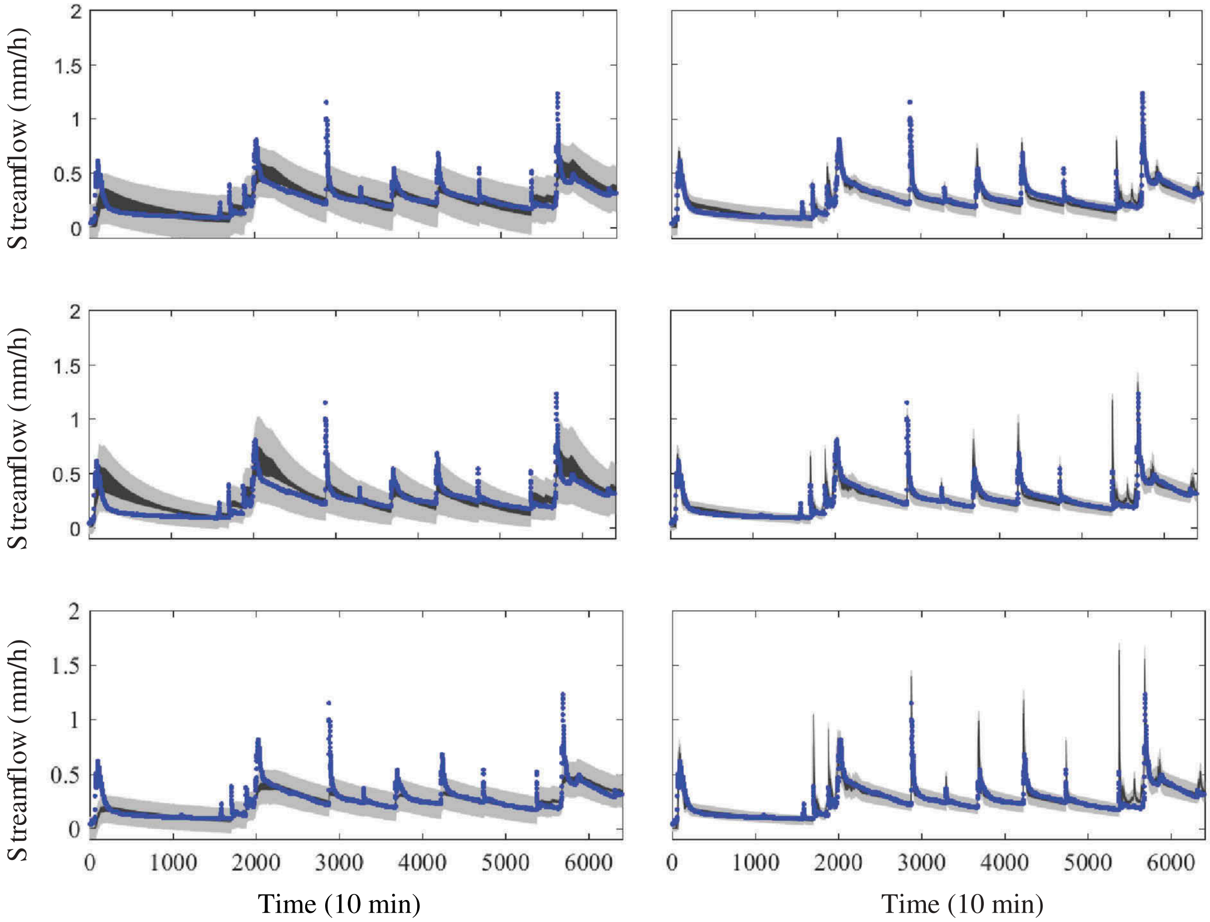


**Figure 7.** Observed runoff series (blue), 95% uncertainty (light grey) and uncertainty associated with the values of the parameters (dark grey) for the Saci River basin for models M03 (left) and M11 (right), using the error models L1 (top), L2 (middle) and L3 (bottom). Calibration was performed with runoff data for the complete series with thinning equal to 72.

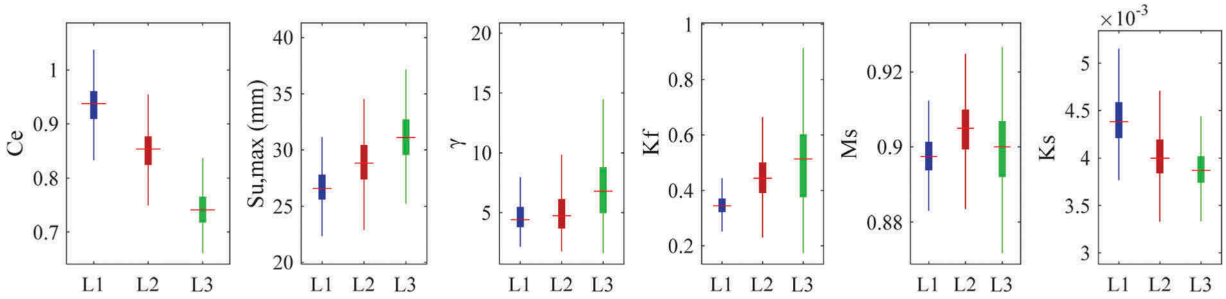


**Figure 8.** Distribution of the parameters inferred with each of the three error models – L1, L2 and L3 – for model M11 in the Saci River basin. Calibration performed with complete runoff data (10 min) with thinning of 72. The central mark indicates the medians, the box indicates the 50% quantiles and the whiskers extend to the 95% quantiles.

selection results obtained for the Saci River basin (based on information criteria only).

For both basins, the models in series presented larger $I_k$ values than the models in parallel (Fig. 9). The model that presented the largest $I_k$ value (i.e. worst) was M03, which is one of the simplest models.

In the validation period, performance of the models for the Bugres River basin considering the three metrics (precision, reliability and volumetric bias metrics) was worse compared to the calibration results (Fig. 10). As in the calibration period, models with a parallel structure resulted in a better performance than models with a serial structure. To illustrate this

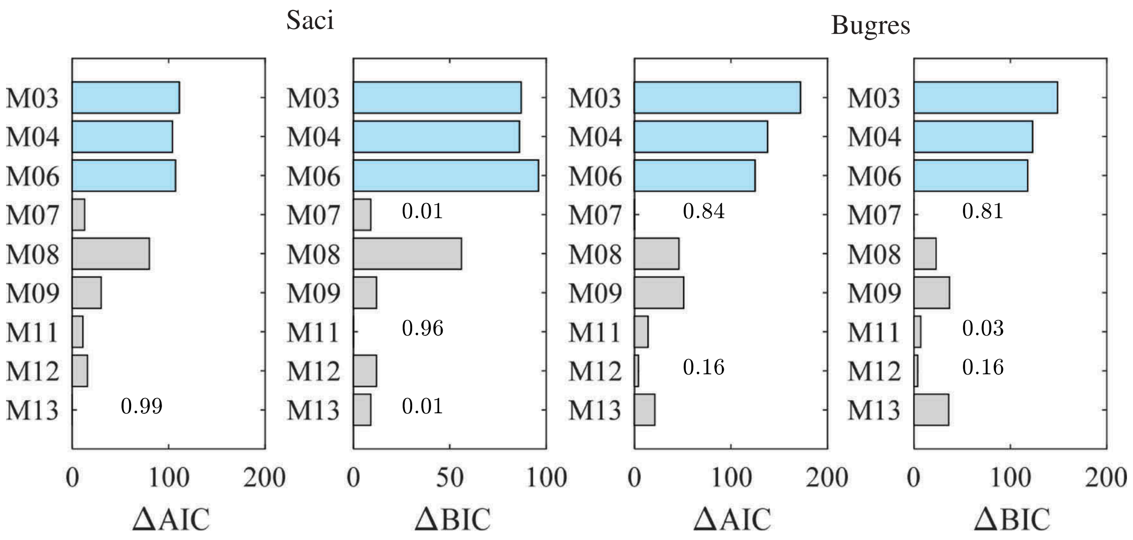
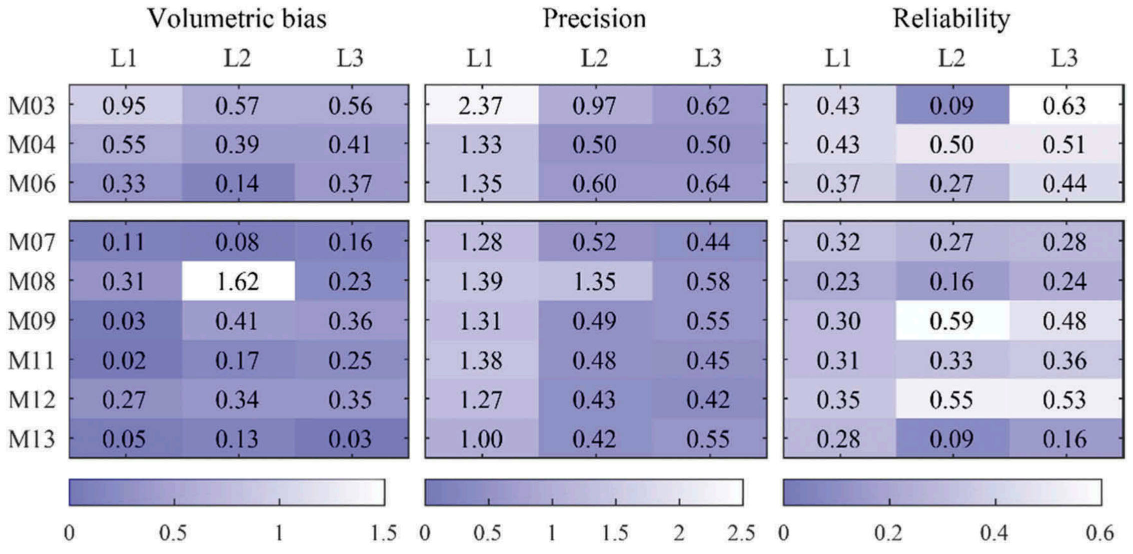


Figure 9. Information criteria results for the Saci and Bugres river basins with error model L3. ΔAIC and ΔBIC correspond the relative support of a model in relation to the best model (the lower the better) for the Akaike information criterion and Bayesian information criterion, respectively. The weights assigned for each model are presented in the text. Weights smaller than 0.01 were not presented. The models in series are in blue and those in parallel are in grey.



| | Volumetric bias | | | Precision | | | Reliability | | |
|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L1 | L2 | L3 | L1 | L2 | L3 |
| M03 | 0.95 | 0.57 | 0.56 | 2.37 | 0.97 | 0.62 | 0.43 | 0.09 | 0.63 |
| M04 | 0.55 | 0.39 | 0.41 | 1.33 | 0.50 | 0.50 | 0.43 | 0.50 | 0.51 |
| M06 | 0.33 | 0.14 | 0.37 | 1.35 | 0.60 | 0.64 | 0.37 | 0.27 | 0.44 |
| M07 | 0.11 | 0.08 | 0.16 | 1.28 | 0.52 | 0.44 | 0.32 | 0.27 | 0.28 |
| M08 | 0.31 | 1.62 | 0.23 | 1.39 | 1.35 | 0.58 | 0.23 | 0.16 | 0.24 |
| M09 | 0.03 | 0.41 | 0.36 | 1.31 | 0.49 | 0.55 | 0.30 | 0.59 | 0.48 |
| M11 | 0.02 | 0.17 | 0.25 | 1.38 | 0.48 | 0.45 | 0.31 | 0.33 | 0.36 |
| M12 | 0.27 | 0.34 | 0.35 | 1.27 | 0.43 | 0.42 | 0.35 | 0.55 | 0.53 |
| M13 | 0.05 | 0.13 | 0.03 | 1.00 | 0.42 | 0.55 | 0.28 | 0.09 | 0.16 |

Figure 10. Performance metrics for the Bugres River basin, using error models L1, L2, and L3. Validation performed with thinning of 72.

result, Figure 11 presents the predictive uncertainty for models M04 and M07 obtained with the different error models. The models with better results considering the three performance metrics together were M07 and M13. Both models are among the most complex models; therefore, it can be said that this complexity is justified.

When comparing the maximum log-likelihood values from validation with the AIC and BIC values, it can be observed that models with lower information criteria values presented better performance in validation (Fig. 12). This result indicates that the use of information criteria values alone for the Saci River basin may be reliable, since both methods lead to the same results.

For the Saci River basin, the models in parallel in general presented smaller values of $I_k$ (Fig. 9). The weights for this model were very close to 1 and models M11 and M12 presented very small weights, indicating that the probability of choosing other models rather than M13 in a different dataset from that used in the calibration is practically zero. Considering the BIC values, the best model was M11. The BIC criterion selected a less complex model than AIC, which indicates that the former criterion penalizes the complexity more than the latter.

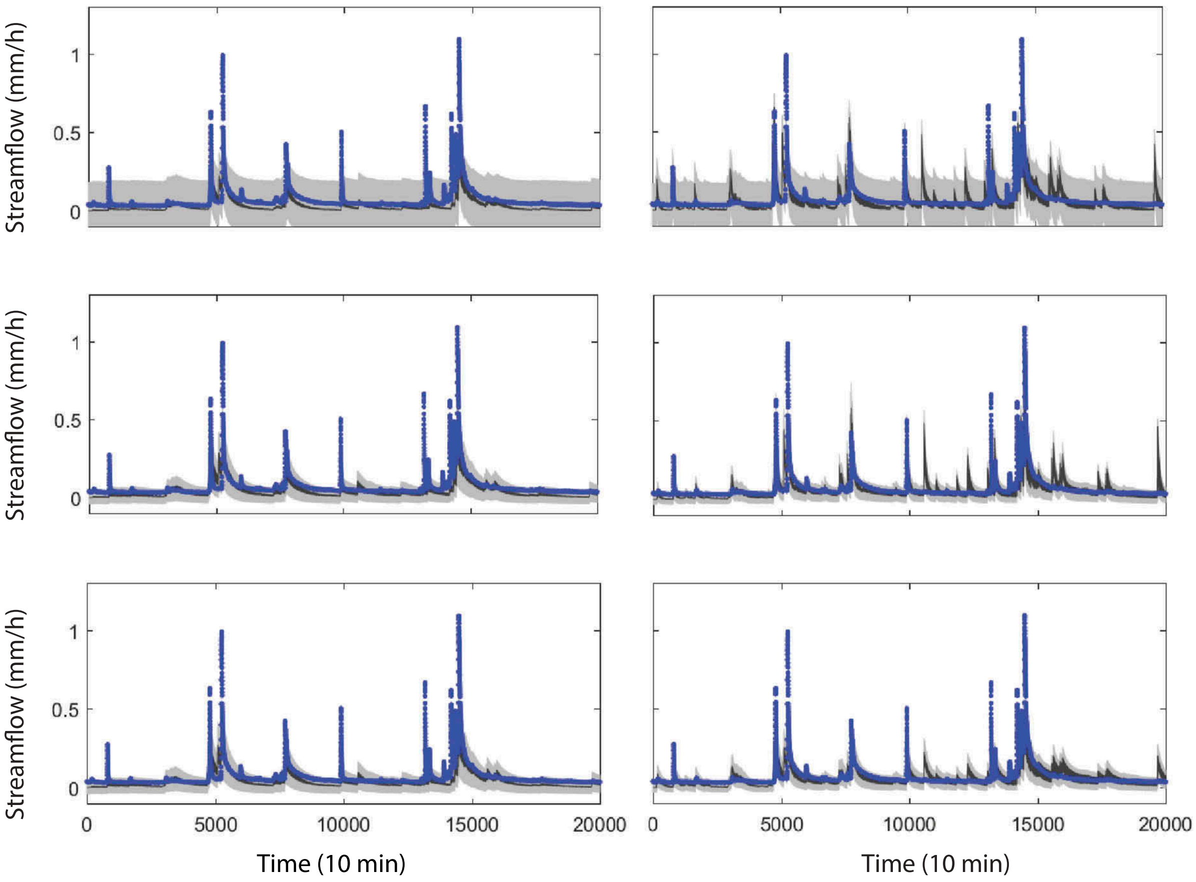


**Figure 11.** Observed runoff series (blue dots), 95% uncertainty (light grey) and uncertainty associated with the values of the parameters (dark grey) for the Bugres River basin for models M04 (left) and M07 (right), using the error models L1 (top), L2 (middle) and L3 (bottom). Validation performed using thinning of 72.

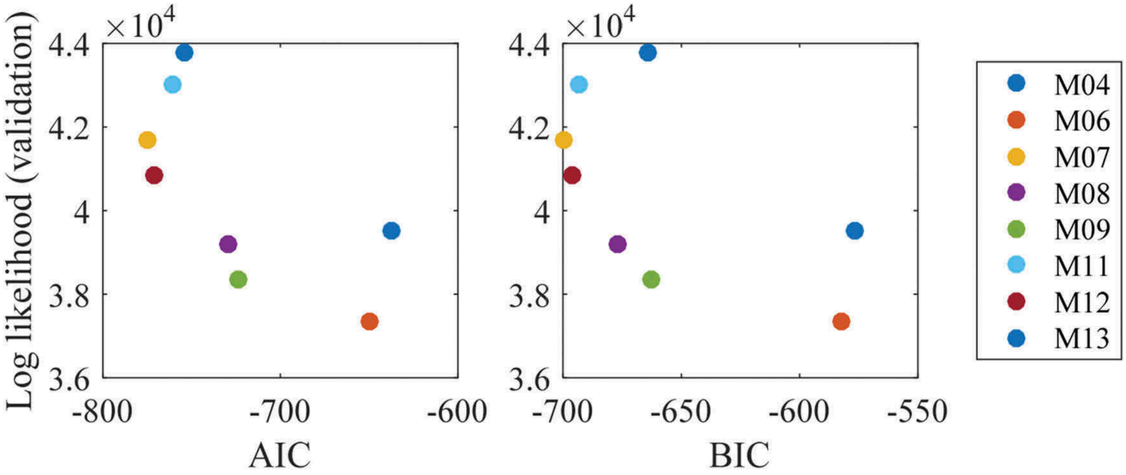


**Figure 12.** Maximum log likelihood values in the validation period for the Bugres River basin against the AIC (left) and BIC (right) values. The validation of M03 is omitted because it presented a discrepant result that made the visualization difficult.

With respect to the different error models, L3 presented smaller (better) information criteria values than the other two for both basins (results not shown), which, along with the results of performance metrics presented before, supports the use of a more complex error model. Also, the rank of the hydrological models obtained using information criteria values was influenced by the choice of the error model used for parameter inference. This result highlights the importance of the correct choice of the error assumptions in order to find the best models.

## Residual diagnostics

The verification of the assumptions about the model residuals was done graphically for each model. All the hydrological models were calibrated with the three error models considered in this study (L1, L2 and L3). The residuals were standardized by dividing them by $\sigma_t$. The standardized residuals were evaluated in relation to their adjustment to the assumed distribution, their variance as a function of the observed runoff value and their temporal autocorrelation. In Figure 13 we show an example for the Bugres River basin. The premises for the residual models were not met in any case for the L1 function; that is, the errors are heteroscedastic, do not follow a normal distribution and are highly correlated. With the L2, which considers the heteroscedasticity of the errors, the error variance was not yet constant according to the observed runoff values and the residuals were correlated. However, a better adjustment to the assumed distribution was obtained. The L3 error model presented similar results for the variance and autocorrelation, and the distribution was closer to the one assumed in the calibration.

Consideration of thinning of the data series increased the spread of the parameter posterior distribution, without significantly changing the median parameter values (Fig. 14).

We also investigated the relation between error parameters and model structures. We plotted the posterior distributions of the error model parameters ($\sigma_0$, $\sigma_1$ and $\beta$) for each hydrological model sorted from best to worst according to the information criterion AIC (Fig. 15). It is possible to verify that models with similar results also have similar error characteristics. Models with lower values of $\sigma_0$ and $\sigma_1$ resulted in a narrower predictive uncertainty and therefore were also the ones with smaller (better) values of the precision metric. The value of the heteroscedasticity intercept ($\sigma_0$) presented larger variations between models in parallel and in series in the Saci River basin than in the Bugres River basin. Models in parallel, which presented better AIC values, resulted in

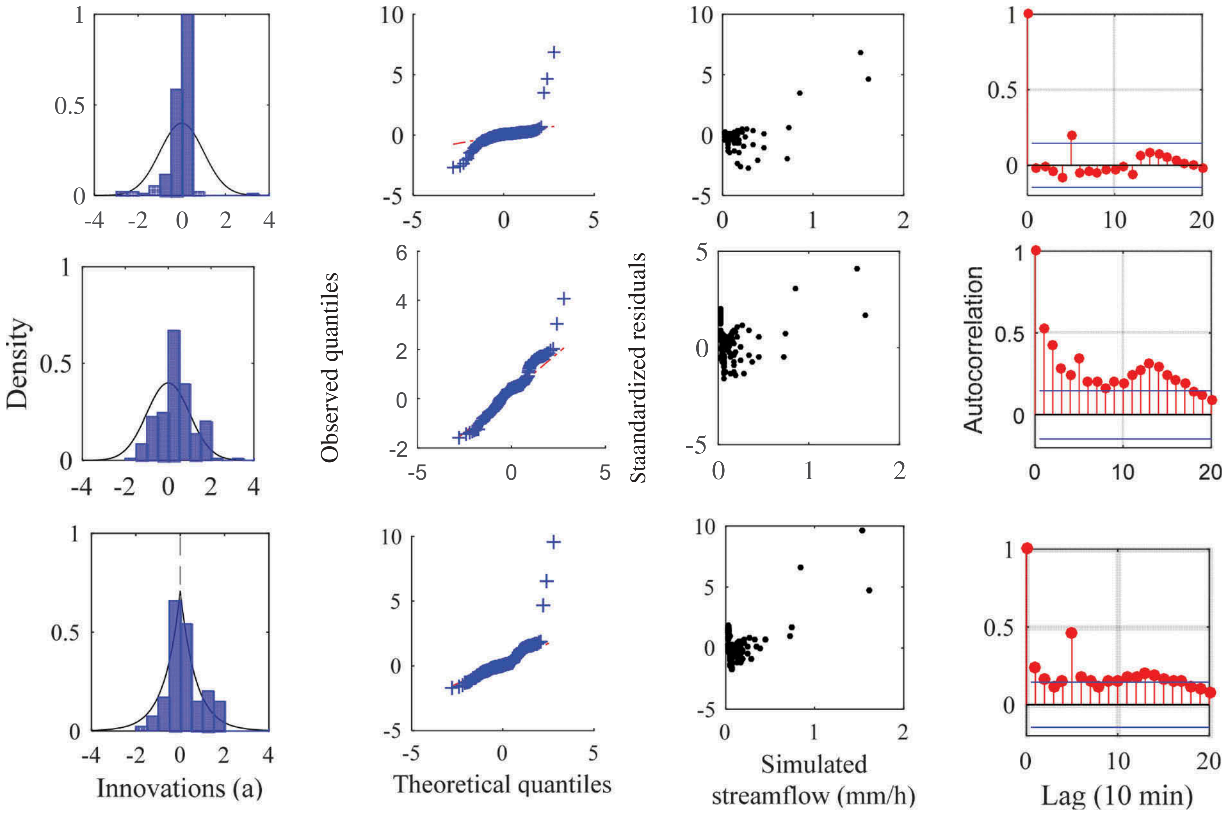


Figure 13. Diagnostics of the residuals for the Bugres River basin. Calibration of model M11 performed with L1 (top), L2 (middle) and L3 (bottom) with temporal resolution of 10 min and thinning of 72. From left to right: (a) histograms of the standardized residuals, where the black line indicates the theoretical distribution; (b) quantile–quantile plot of standardized residuals and the theoretical distribution; (c) standardized residuals as a function of simulated streamflow; and (d) the autocorrelation function of the standardized residuals, where the blue lines indicate the 95% significance levels.

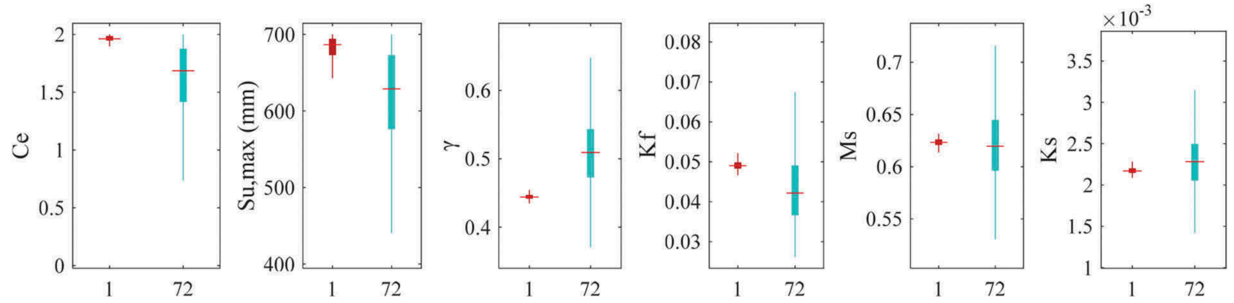


Figure 14. Posterior distribution of the parameters of model M11 for the Bugres River basin, using the error model L3. Calibration performed with thinning of 1 (red) and 72 (green). The central mark indicates the medians, the box indicates the 50% quantiles and the whiskers extend to the 95% quantiles.

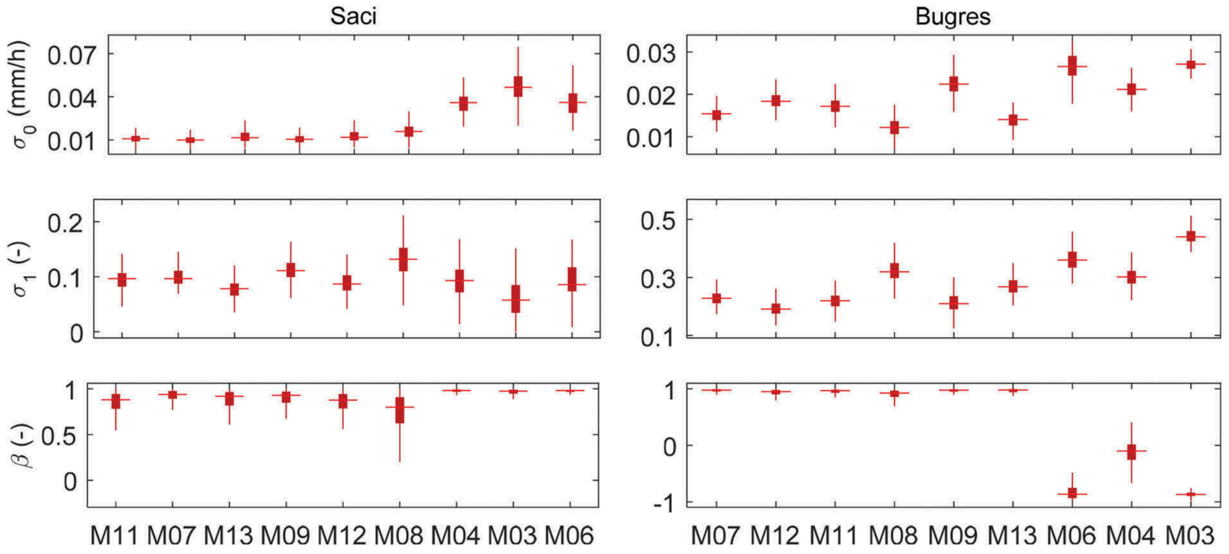


Figure 15. Posterior distribution of the parameters of the error model L3 –heteroscedasticity intercept $\sigma_0$, heteroscedasticity slope $\sigma_1$, and kurtosis $\beta$ – for the Saci and Bugres river basins. The hydrological models are ordered from best to worst according to the information criteria AIC. The central mark indicates the medians, the box indicates the 50% quantiles, and the whiskers extend to the 95% quantiles.

smaller values of $\sigma_1$; therefore, a lower degree of heteroscedasticity.

For the Saci River basin, the kurtosis parameter ($\beta$) converged to the value of 1, meaning that the errors follow a Laplace distribution, independently of the model used. This kind of distribution has heavier tails, which makes it robust against outliers. For the Bugres River basin, there is a larger difference between $\beta$ values for models in parallel and in series. For models in series, which resulted in the worst AIC values, $\beta$ was closer to −1, representing a uniform distribution.

## Discussion

### *Model structure informing about dominant runoff generation processes*

Consideration of a parallel structure and of the non-linearity of the unsaturated zone better simulated the catchment outflow of the Bugres River basin. The results from this study site are in agreement with the knowledge of its hydrological functioning (Grison *et al.* 2014), characterized by a rapid response to rainfall (represented by the fast reservoir) and also a constant river discharge fed by water storage in the thick soil layer.

The models with a parallel structure also had significantly better performance than those in series for the Saci River basin. In addition to the parallel structure, the presence of an unsaturated soil reservoir was also of major importance. This result indicates that, for this catchment, the soil exerts an important role in the release and storage of water. This basin has a hydrologically active layer (with weathered material) of more than 5 m depth in most of its area, ranging from 6 m to less than 1 m near the river spring (Santos 2009). Santos (2009) verified that the soil in the basin has a high infiltration capacity and that even with high precipitation intensities this capacity was not exceeded. Those characteristics justify the importance of the unsaturated soil reservoir and the division of the flow between fast and slow reservoirs in this catchment.

Even though the two basins are forested, the inclusion of an interception reservoir did not improve model performance. That insensitivity to explicitly modelling interception may be due to the fact that the models are lumped and consider that the whole basin behaves in the same way, without considering the spatial distribution of rain and vegetation. Moreover, for the Bugres River basin the $C_e$ parameter – an adjustment for the evapotranspiration – converged to 2 for all models, regardless of the consideration of the interception process. This result might be an attempt to compensate for other losses not explicitly accounted for by the model. In the case of the Saci River basin, the $C_e$ parameter was close to 1 for all models.

### Is the increase in model complexity justified?

The information criteria considered in this study tend to favour an increase in model complexity in both basins but did not lead to the selection of the most complex model (in terms of the number of parameters). Westra *et al.* (2014) also found similar results when using the AIC for model selection. The variation in the term related to the maximum likelihood function value was much greater than the variation in the term that penalizes complexity. Therefore, the model that produced a higher maximum likelihood function value also resulted in better information criteria values. For the Bugres River basin, the best models evaluated with AIC and BIC in the calibration period also performed better in the validation period, i.e. lower AIC and BIC values in calibration were associated with higher maximum log likelihood values in validation. Similar results were found by Westra *et al.* (2014) with AIC. In their work, even though the AIC favoured more complex models, these models performed better in the validation period. For the Saci River basin the AIC and BIC led to different results. The BIC criterion penalized the complexity more than the AIC and, therefore, selected a model less complex (in terms of the number of parameters) than the AIC criterion (M11, with seven parameters, *versus* M13, with eight parameters). Furthermore, we highlight the importance of the correct choice of the error assumptions in order to correctly identify the best models, since both AIC and BIC were sensitive to the error model considered.

Models with the same number of parameters, such as M04 and M09 (with five each), had very different performance metrics and information criteria values. Those results, together with the fact that the most complex model (M12, with nine parameters) was not considered the best, are evidence that the model structure is more important than the number of parameters.

### Model structure and residual error parameters

The residual errors of hydrological models were best represented by a model that explicitly accounts for heteroscedasticity, since larger streamflow values are often associated with larger error measurements, and non-normality. In addition, a wrong representation of the residuals resulted in a poorer predictive uncertainty, as found in other studies (e.g. Kavetski *et al.* 2011).

The posterior distribution of the error model parameters for the different hydrological models shows that a model structure that better represents the simulated discharge results in a lower degree of heteroscedasticity, as evidenced by the lower values of the heteroscedasticity slope ($\sigma_1$). The heteroscedasticity intercept ($\sigma_0$) represents the standard deviation of residuals for low flows. In the Saci River basin there was a clear difference in the simulation of low flows between serial and parallel structures. In the Bugres River basin, $\sigma_0$ values did not vary so much between parallel and serial structures, and the simulation of low flows between those structures was not so different as in the Saci.

## Summary and conclusions

In this study we combined flexible hydrological modelling, uncertainty analysis and control of model complexity to identify model structures that have a better correspondence with catchment behaviour and hence hypothesize about dominant runoff generation mechanisms. We compared the performance of three error models and nine conceptual hydrological models from the SUPERFLEX framework. Model performance was evaluated considering both information criteria, which penalize for increasing model complexity, and the quality of streamflow predictive uncertainty. As a study case, the proposed methodology was applied to the rainfall–runoff modelling of two forested basins located in the southern region of Brazil.

The use of model structures that differed systematically from each other allowed verification of the impact of different components on the results, such as the inclusion of reservoirs and nonlinearity of flows. The main difference in performance considering both the quality of the predictive uncertainty (measured by the performance metrics) and the information criteria values resulted from the consideration of a parallel connection between the fast and slow reservoirs. This result indicates that the model architecture was more important than the increase in the number of model parameters in the representation of the runoff generation in these two basins.

With the increase in error model complexity, residuals assumptions were better satisfied, and the quality of the predictive uncertainty was improved. In addition, each error model led to different parameter posterior distributions. Therefore, we highlight the importance of the proper choice of the error model since it affected the quality of the predictive uncertainty, the inferred parameter values, and also the model selection with information criteria. The posterior distribution of the error model parameters for the different hydrological models showed that a model structure that better represents the simulated discharge results in a lower degree of heteroscedasticity, as evidenced by the lower values of the heteroscedasticity slope.

Considering the quality of the predictive uncertainty, information criteria and model performance in the validation period, the same models presented better results, showing that the method utilized in this study was consistent. The information criteria were inversely related with the maximum log likelihood values from the validation period, meaning that even though more complex models were favoured by the selected information criteria, this complexity was justified by the data. Nevertheless, we believe that more robust results would be found with the use of a model selection method that considers the values of the likelihood function over the entire parameter space, and not only the maximum value of the likelihood function. An example would be to use the evidence, i.e. the denominator of the Bayes theorem, as suggested by Volpi *et al.* (2017).

More robust conclusions could be made if additional observations were available, as done by some recent studies. McMillan *et al.* (2012) utilized tracer dynamics as a diagnostic tool to evaluate model structures. Kuppel *et al.* (2018) combined a physically-based model with water isotopic tracer and age. Knighton *et al.* (2017) utilized water isotope tracers to verify the assumptions about the unsaturated zone. Despite the use of a relatively short data series, especially as was the case for the Saci River basin, the rigorous methodology presented here allowed the acquisition of some insights about the behaviour of the studied catchments. This methodology can be very useful when types of data other than precipitation and streamflow are not available for model constraint.

## Acknowledgments

The authors gratefully acknowledge the Associate Editor of *Hydrological Sciences Journal* and Mr. Knighton and the other two anonymous reviewers whose suggestions greatly improved the manuscript.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and Conselho Nacional de Desenvolvimento Científico e Tecnológico and through scholarship for the first, second and fourth authors.

## ORCID

Paula C. David http://orcid.org/0000-0003-1627-7461
Debora Y. Oliveira http://orcid.org/0000-0003-3635-3249
Fernando Grison http://orcid.org/0000-0002-5256-8744
Masato Kobiyama http://orcid.org/0000-0003-0615-9867
Pedro L. B. Chaffe http://orcid.org/0000-0002-9918-7586

## References

Akaike, H., 1974. *A new look at the statistical model identification. IEEE Transactions on Automation Control*, 19 (6), 716–723.

Boer-Euser, T., *et al.*, 2017. Looking beyond general metrics for model comparison - Lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, 21, 423–440. doi:10.5194/hess-21-423-2017

Butts, M.B., *et al.*, 2004. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology*, 298, 242–266. doi:10.1016/j.jhydrol.2004.03.042

Chaffe, P.L.B., *et al.*, 2010. Is interception information important for rainfall-runoff modeling? *Annual Journal of Hydraulic Engineering JSCE*, 54, 91–96.

Clark, M.P., *et al.*, 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44, 1–14. doi:10.1029/2007WR006735

Doorenbos, J. and Pruitt, W.O., 1977. *Crop water requirement*. Rome: Food and Agricultural Organization of the United Nations, 144.

Evin, G., *et al.*, 2013. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resources Research*, 49 (7), 4518–4524. doi:10.1002/wrcr.20284

Evin, G., *et al.*, 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, 50, 2350–2375. doi:10.1002/2013WR014185

Fenicia, F., et al., 2006. Is the groundwater reservoir linear? Learning from data in hydrological modelling. *Hydrology and Earth System Sciences*, 10, 139–150. doi:10.5194/hess-10-139-332002006

Fenicia, F., *et al*., 2014. Catchment properties, function, and conceptual model representation. Is there a correspondence? *Hydrological Processes*, 28, 2451–2467. doi:10.1002/hyp.9726

Fenicia, F., Kavetski, D., and Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, 47, 1–13. doi:10.1029/2010WR010174

Fenicia, F., McDonnell, J.J., and Savenije, H.H.G., 2008. Learning from model improvement: on the contribution of complementary data to process understanding. *Water Resources Research*, 44, 1–13. doi:10.1029/2007WR006386

Gao, H., *et al*., 2014. Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. *Hydrology and Earth System Sciences*, 18, 1895–1915. doi:10.5194/hess-18-1895-2014

Grison, F., Mota, A., and Kobiyama, M., 2014. Geometria hidráulica de seções transversais do rio dos Bugres. *Revista Brasiliera De Recursos Hídricos (Brazilian Journal of Water Resources)*, 19, 205–213. doi:10.21168/rbrh.v19n4.p205-213

Kavetski, D. and Fenicia, F., 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and Experimental Insights. *Water Resources Research*, 47, 1–19. doi:10.1029/2011WR010748

Kavetski, D., Fenicia, F., and Clark, M.P., 2011. Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: insights from an experimental catchment. *Water Resources Research*, 47, 1–25. doi:10.1029/2010WR009525

Knighton, J.O., DeGaetano, A., and Walter, M.T., 2017. Hydrologic state influence on riverine flood discharge for a small temperate watershed (Fall Creek, United States): negative feedbacks on the effects of climate change. *Journal of Hydrometeorology*, 18 (2), 431–449. doi:10.1175/JHM-D-16-0164.1

Kuppel, S., *et al*., 2018. EcH2O-iso 1.0: water isotopes and age tracking in a process-based, distributed ecohydrological model. *Geoscientific Model Development*, 11 (7), 3045–3069. doi:10.5194/gmd-11-3045-2018

Laloy, E. and Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. *Water Resources Research*, 48, 1–18. doi:10.1029/2011WR010608

Lever, J., Krzywinski, M., and Altman, N., 2016. Points of Significance: model selection and overfitting. *Nature Methods*, 13, 703–704. doi:10.1038/nmeth.3968

Li, H., Xu, C.Y., and Beldring, S., 2015. How much can we gain with increasing model complexity with the same model concepts? *Journal of Hydrology*, 527, 858–871. doi:10.1016/j.jhydrol.2015.05.044

McGuire, K.J. and McDonnell, J.J., 2010. Hydrological connectivity of hillslopes and streams: characteristic time scales and nonlinearities. *Water Resources Research*, 46, 1–17. doi:10.1029/2010WR009341

McInerney, D., *et al*., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, 53, 2199–2239. doi:10.1002/2016WR019168

McMillan, H., *et al*., 2012. Do time-variable tracers aid the evaluation of hydrological model structure? A multimodel approach. *Water Resources Research*, 48, 5. doi:10.1029/2011WR011688

Oliveira, D.Y., Chaffe, P.L.B., and Sá, J.H.M., 2018. Extending the applicability of the generalized likelihood function for zero-inflated data series. *Water Resources Research*, 54, 2494–2506. doi:10.1002/2017WR021560

Orth, R., *et al*., 2015. Does model performance improve with complexity? A case study with three hydrological models. *Journal of Hydrology*, 523, 147–159. doi:10.1016/j.jhydrol.2015.01.044

Perrin, C., Michel, C., and Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242, 275–301. doi:10.1016/S0022-1694(00)00393-0

Poncelet, C., *et al*., 2017. Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. *Water Resources Research*, 53, 2742–2759. doi:10.1002/2016WR019991

Santos, I., 2009. *Monitoramento e modelagem de processos hidrogeomorfológicos: Mecanismos de geração de escoamento e conectividade hidrológica*. Thesis (PhD). Universidade Federal de Santa Catarina, Brazil.

Savenije, H.H.G. and Hrachowitz, M., 2017. HESS opinions "Catchments as meta-organisms - A new blueprint for hydrological modelling.". *Hydrology and Earth System Sciences*, 21, 1107–1116. doi:10.5194/hess-21-1107-2017

Schöniger, A., *et al*., 2014. Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50, 9484–9513. doi:10.1002/2014WR016062

Schoups, G., Van De Giesen, N.C., and Savenije, H.H.G., 2008. Model complexity control for hydrologic prediction. *Water Resources Research*, 44, 1–14. doi:10.1029/2008WR006836

Schoups, G., Vrugt, J. A., Fenicia, F., and van de Giesen, N. C., 2010. Corruption of accuracy and efficiency of markov chain monte carlo simulation by inaccurate numerical implementation of conceptual hydrologic models. *Water Resources Research*, 46, 1–12. doi: 10.1029/2009WR008648

Schoups, G. and Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46, 1–17. doi:10.1029/2009WR008933

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461–464. doi:10.1214/aos/1176344136

Smith, T., *et al*., 2010. Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resources Research*, 46, 1–11. doi:10.1029/2010WR009514

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

Sugawara, M., 1961. On the analysis of runoff structure about several Japanese rivers. *Japanese Journal of Geophysics*, 2 (4), 1–76.

Sugawara, M., 1995. Tank Model. *In*: V.P. Singh, ed. *Computer models of watershed hydrology*. Highlands Ranch, CO: Water Resources Publications, 165–214.

Thyer, M., *et al*., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*. 45. doi:10.1029/2008WR006825

Vaché, K.B. and McDonnell, J.J., 2006. A process-based rejectionist framework for evaluating catchment runoff model structure. *Water Resources Research*, 42, 1–15. doi:10.1029/2005WR004247

Van Der Linden, S. and Woo, M.K., 2003. Application of hydrological models with increasing complexity to subarctic catchments. *Journal of Hydrology*, 270, 145–157. doi:10.1016/S0022-1694(02)00291-3

Van Esse, W.R., *et al.*, 2013. The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. *Hydrology and Earth System Sciences*, 17, 4227–4239. doi:10.5194/hess-17-4227-2013

Volpi, E., Schoups, G., Firmani, G., and Vrugt, J. A., 2017. Sworn testimony of the model evidence: gaussian mixture importance (game) sampling. *Water Resources Research*, 53, 6133–6158. doi: 10.1002/2016WR020167

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. *Environmental Modelling and Software*, 75, 273–316. doi:10.1016/j.envsoft.2015.08.013

Westra, S., *et al.*, 2014. A strategy for diagnosing and interpreting hydrologicalmodel nonstationarity. *Water Resources Research*, 1–24. doi:10.1002/2013WR014719