



## Bayesian joint inference of hydrological and generalized error models with the enforcement of Total Laws

Mario R. Hernández-López<sup>1</sup>, Félix Francés<sup>1</sup>

<sup>1</sup> Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, Spain

Correspondence to: Mario R. ([maherlo@hma.upv.es](mailto:maherlo@hma.upv.es))

### Abstract.

Over the years, the Standard Least Squares (SLS) has been the most commonly adopted criterion for the calibration of hydrological models, despite the fact that they generally do not fulfill the assumptions made by the SLS method: very often errors are autocorrelated, heteroscedastic, biased and/or non-Gaussian. Similarly to recent papers, which suggest more appropriate models for the errors in hydrological modeling, this paper addresses the challenging problem of jointly estimate hydrological and error model parameters (joint inference) in a Bayesian framework, trying to solve some of the problems found in previous related researches. This paper performs a Bayesian joint inference through the application of different inference models, as the known SLS or WLS and the new GL++ and GL++Bias error models. These inferences were carried out on two lumped hydrological models which were forced with daily hydrometeorological data from a basin of the MOPEX project. The main finding of this paper is that a joint inference, to be statistically correct, must take into account the joint probability distribution of the state variable to be predicted and its deviation from the observations (the errors). Consequently, the relationship between the marginal and conditional distributions of this joint distribution must be taken into account in the inference process. This relation is defined by two general statistical expressions called the Total Laws (TLs): the Total Expectation and the Total Variance Laws. Only simple error models, as SLS, do not explicitly need the TLs implementation. An important consequence of the TLs enforcement is the reduction of the degrees of freedom in the inference problem namely, the reduction of the parameter space dimension. This research demonstrates that non-fulfillment of TLs produces incorrect error and hydrological parameter estimates and unreliable predictive distributions. The target of a (joint) inference must be fulfilling the error model hypotheses rather than to achieve the better fitting to the observations. Consequently, for a given hydrological model, the resulting performance of the prediction, the reliability of its predictive uncertainty, as well as the robustness of the parameter estimates, will be exclusively conditioned by the degree in which errors fulfill the error model hypotheses.



## 1 1. Introduction

2 In hydrologic simulation, many parameters are not directly observable or easily inferred from measured data. For this reason,  
3 these parameters must be estimated indirectly by an inverse process (also called calibration) that conditions the parameter  
4 estimates and the model response on historically observed input-output data (Pokhrel and Gupta, 2010). Model parameter  
5 inferences are based on a likelihood function which quantifies the probability that the observed data were generated by a  
6 particular parameter set (Box and Tiao, 1992). Over the years, the Standard Least Squares (SLS) criterion has been the most  
7 commonly adopted method for the calibration of the unknown parameters in hydrological modeling, although it has long  
8 been known that errors of hydrologic models are generally autocorrelated and heteroscedastic (Sorooshian and Dracup, 1980)  
9 not fulfilling the underlying hypotheses to SLS. Also, model inputs and model structural errors are important sources of  
10 uncertainty and they are ignored by the SLS scheme. I.e., the SLS parameter estimates are adjusted with a particular  
11 calibration data set to partially compensate the input errors (Kavetski et al., 2003; Schoups and Vrugt, 2010) and/or the  
12 structural errors (Schoups and Vrugt, 2010) yielding a set of biased parameters.

13 The present paper is framed as a formal Bayesian aggregated approach to characterize the parameter and total predictive  
14 uncertainty, with a direct method for handling the bias, autocorrelation, heteroscedasticity and non-normality of the  
15 hydrological model errors. This approach follows to that proposed in Schoups and Vrugt (2010), but in present paper we have  
16 introduced significant improvements. Schoups and Vrugt (2010) proposed a formal likelihood function based on a general  
17 error model (called GL error model) that allows for model bias and for autocorrelation, non-stationarity, and non-normality  
18 of model errors. This formal approach preserved the advantages of being developed from a theoretical basis and having the  
19 possibility of checking the assumptions, while at the same time it improved flexibility and reduced the need for unrealistic  
20 assumptions about the errors. That study was based on the joint inference of both the hydrological and error model  
21 parameters (hereinafter joint inference). The main contribution of that research was the treatment of the error  
22 heteroscedasticity and non-normality with a direct method instead of using a transformational one, as it had been previously  
23 used in several case studies (Bates and Campbell, 2001; Kuczera, 1983; Reichert and Mieleitner, 2009; Vrugt et al., 2009b).  
24 However, according to Schoups and Vrugt (2010), something in the method seemed not to work properly: the analysis  
25 performed with heteroscedasticity, autocorrelation, and non-normality error parameters for the second case study (Guadalupe  
26 River basin) gave results with “large and meaningless” prediction uncertainty bands. These authors pointed out that the  
27 reason was the large inferred value for the autocorrelation coefficient. To avoid this problem, the authors decided to adjust  
28 the autocorrelation parameter to an observed sampling value, rather than inferring it jointly and automatically.

29 Evin et al. (2013) described the challenges of fitting hydrological model parameters jointly with the autocorrelation and  
30 heteroscedasticity parameters of error models, and found a solution to the aforementioned problem in Schoups and Vrugt  
31 (2010). The key to the problem solution depended on how the autoregressive model was applied. Applying it directly to the  
32 errors can produce instability in the computations and, as a result, a poor error model with large predictive uncertainty. This  
33 instability could be avoided by applying the autoregressive process to the studentized errors. Thus, the heteroscedasticity  
34 model was performed on errors (instead on innovations) and the de-correlation process was applied on the previously  
35 variance-stabilized errors.

36 In a useful and comprehensive research on the comparison of inference methodologies, Evin et al. (2014) detected that  
37 problems remained in the joint estimation of both hydrological and error model parameters. They concluded that the joint  
38 inference could be non-robust due to multiway interactions between the hydrological parameters (related with the water  
39 balance in their case study) and the error model parameters, particularly, heteroscedasticity and autocorrelation error model  
40 parameters.

41 Moreover, several related researches have avoided a full joint inference either without the modeling of the error’s  
42 autocorrelation (Cheng et al., 2014; Westra et al., 2014) or using a transformational (or indirect) method for the treatment of  
43 the error heteroscedasticity (Bates and Campbell, 2001; Cheng et al., 2014; Del Giudice et al., 2013; Kuczera, 1983; Reichert  
44 and Mieleitner, 2009; Vrugt et al., 2009b). Even there are authors who combine the Markov Chain Monte Carlo (MCMC)  
45 sampling of the hydrological parameters with a maximum likelihood estimation of the error model parameters conditioned on  
46 each MCMC hydrological parameter sample (Zheng and Han, 2015; Han and Zheng, 2016), making a sort of pseudo-joint



47 inference. However other authors as Scharnagl et al.(2015) tried the joint inference without success, finding similar problems  
48 to those exposed in Schoups and Vrugt (2010) or Evin et al. (2013, 2014).

49 Consequently, the general purpose of our research is to find out the reasons and to overcome the drawbacks identified in the  
50 previous studies, defending the reliability of the Bayesian joint inference methodology. This defense is necessary because  
51 non-using the joint inference involves a range of problems (Evin et al., 2014): i) the use of a faulty error model when  
52 calibrating the hydrological parameters produces biased parameter estimates and an incorrect parameter and predictive  
53 uncertainty estimates; ii) the estimation of error model parameters conditioned on a particular estimate of hydrological  
54 parameters (assuming that they could be biased) ignores interactions between both sets of parameters, and may produce an  
55 incorrect estimation of the predictive uncertainty.

56 The specific objectives covered by this paper are the following. Firstly, we have improved the formal general likelihood  
57 function for parameter and predictive inference of hydrological models developed by Schoups and Vrugt (2010),  
58 implementing the previously mentioned recommendation proposed in Evin et al. (2013). Secondly, this paper seeks to shed  
59 some light on the causes of the incorrect behavior of the joint inference which has been detected in Schoups and Vrugt  
60 (2010), Evin et al. (2013, 2014) and Scharnagl et al. (2015).

61 Todini (2007) pointed out that “(...) *uncertainty (...) is 'conditional' upon the forecasting model prediction*”. That idea is the  
62 basis to understand that the concept of predictive uncertainty must be linked with the conditional distribution of a predicted  
63 variable given its related model prediction namely  $p(y|y_s)$ . Therefore, the existence of that conditional distribution  
64 implies also the existence of its joint distribution  $p(y, y_s)$ , which is the foundation of the general framework proposed in  
65 this paper. When we define the conditional distributions through any of its statistical features (i.e. shape, position, etc.), it is  
66 important to note that we are defining in parts their parent joint distribution. So, these parts must comply with the necessary  
67 conditions/restrictions to conform all together such joint distribution. In particular, two necessary conditions are the Total  
68 Variance Law and the Total Expectation Law (for details about these laws see for example Blitzstein and Hwang, 2014),  
69 which relate the marginal and conditional distributions of a given joint distribution. This paper enforces both Total Laws to  
70 the heteroscedasticity and non-constant bias models, since they are part of the conditional distributions modeling and, hence,  
71 they belong to the joint probability density function (hereinafter pdf). The presented case studies will show how not enforcing  
72 the Total Laws could be the origin of the previously found problems with the Bayesian joint inference.

73 The next section presents the generalized error statistical model followed by this research. “General” in the sense that error  
74 model permits the possibility for the errors to be autocorrelated, heteroscedastic, biased and/or non-Gaussian. Section 3 deals  
75 with the joint Bayesian parameter inference. A new formal generalized likelihood function is presented and the methods for  
76 obtaining the posterior of parameters and the predictive distribution are outlined. Section 4 describes the different inference  
77 settings and how to apply the Total Laws on them. Section 5 applies the exposed methodology to estimate the parameter and  
78 predictive uncertainty, through the joint inference with several error models and two lumped hydrological daily models, using  
79 hydrometeorological data from a basin in the Model Parameter Estimation Experiment (MOPEX) data set (Duan et al., 2006).  
80 The fulfillment of assumptions in the hypothesized error models, as well as several indicators of performance will be tested.  
81 Section 6 discusses the results and section 7 summarizes our findings.

## 82 **2. Generalized error statistical model**

### 83 **2.1 Why do we need a reliable error model?**

84 Let us consider a random variable of hydrological predictive interest to be forecasted, for which the observations are made  
85 and which will be called the predictand  $y$  (e.g. the streamflow at a catchment outlet  $q$ , or any other state variable of  
86 interest). The predictand  $y$  can be sampled at  $N$  constant time steps and it is jointly observed and sampled with a set of  
87 predictors or model inputs,  $\mathbf{X}$  (e.g. precipitation, temperature, etc.). This research will consider  $\mathbf{X}$  as deterministic  
88 variables, which are able to explain “something” about the predictand’s behavior.



89 Let us also consider any model output  $y_s$  (e.g. the simulated streamflow  $q_s$ , or some other observable and simulated state  
90 variable) as a random variable; not because a stochastic model is used (which is not our case), but by the fact that we can  
91 observe that these simulated variables generally do not match their observations. In this sense, Todini (2007) pointed out that  
92 “ (...) a scatter will always be observed in the  $q$ - $q_s$  plane (...) a representation of the joint sample frequency of  $q$  and  $q_s$  that  
93 can be used to estimate its joint probability density”. That is to say, there is an inherent uncertainty whose origin is in the  
94 model structure soundness, in the observed data errors of predictand and predictors, and ultimately in the inherent  
95 unpredictability in deterministic terms of the natural phenomena. Montanari and Koutsoyiannis (2012) concluded that  
96 uncertainty is unavoidable in Hydrology so it is impossible to produce a fully deterministic model that would eliminate the  
97 uncertainty, and modeling schemes need to explicitly recognize its role.

98 As stated in Todini (2007), two different objectives can be differentiated in hydrological modeling: parameter estimation and  
99 hydrological prediction. Parameter estimation is the procedure for obtaining parameter values with a physical meaning, which  
100 help us to understand the nature of the modeled processes. That is to say, the aim of the hydrological parameter estimation is  
101 not (it should not be) to get the best fit with the observations, but achieving the most plausible estimates with the best  
102 possible error model for the given observed data set. Hydrologists should assume the fact that, given a correct error model  
103 (actually a correct likelihood function), the most plausible parameter set can produce simulated results which show a poor fit  
104 to the observations, even in the calibration period. In other words, an inference with the correct error model yields the most  
105 plausible parameter set, which will not necessarily yield the best fit to the observed data, but will produce the best possible  
106 result, taking into account the limitations of the hydrological model and the observational errors. In a classical Frequentist  
107 context, the most plausible parameters provide the Maximum Likelihood and in a Bayesian framework are the Maximum a  
108 Posteriori (from now MAP).

109 With regard to the problem of making predictions about some variable of interest, the correct estimation of the hydrological  
110 parameters is not necessary, since the predictive uncertainty,  $p(y|y_s)$ , can be obtained through the direct modeling of the  
111 joint distribution of the predictand and its related model output random variable, namely the  $p(y, y_s)$ , as it was first shown  
112 by Krzysztofowicz (1999) and many others later. Montanari and Brath (2004), under certain assumptions, carried out the  
113 predictive uncertainty assessment through the modeling of the joint distribution of the model errors and the model predictions  
114  $p(e, y_s)$  instead of modeling  $p(y, y_s)$ , although all of them formulated the problem by using a meta-Gaussian model  
115 (Kelly and Krzysztofowicz, 1997).

116 When we want to meet both objectives parameter estimation and predictive uncertainty assessment at the same time, we can  
117 address the issue through the modeling of the arising uncertainty from the uncertainty sources. For this task we can use two  
118 methods: the aggregate method or the disaggregate approach (Kavetski et al., 2006a, 2006b; Kuczera et al., 2006; Vrugt et  
119 al., 2008; Reichert and Mieleitner, 2009; Renard et al., 2010, 2011). The aggregate method, on which this research is focused,  
120 deals with modeling the hydrological model errors, considering the aggregated effects on them (i.e. bias, autocorrelation,  
121 heteroscedasticity, etc.), which are produced by all the uncertainty sources. This is made without the need to refer the  
122 particular contribution of each uncertainty source to these effects. If we are able to pack all the knowledge about the  
123 deviations of the model results from the observations, in an appropriate likelihood function (actually an error model), and the  
124 observed data meet the system observability conditions (Gelb, 1974), only then it will be possible a correct hydrological  
125 parameter estimation which yields unbiased, accurate and physically meaningful parameter values (Sorooshian and Gupta,  
126 1983). This estimation necessarily needs to be a joint inference of error and hydrological models in order to avoid biased  
127 parameters (Evin et al., 2014). Moreover, since the modeling of the joint distribution  $p(y, y_s)$  can be equivalent, under  
128 certain assumptions which will be exposed in the following section 2.2, to the modeling of the joint distribution of the errors  
129 and the model predictions  $p(e, y_s)$ , then a correct error model will also yield correct predictive uncertainty estimation. So,  
130 the answer to the question posed in the title of this section is that the closer we get to the ideal conditions for a correct error  
131 modeling, which may be a difficult task in hydrology (Smith et al., 2015), the more exact and accurate will be both the  
132 parameter estimation and the predictive uncertainty evaluation.



133 **2.2 The error model**

134 The relation between a predictand  $y$  and a model prediction  $y_s$  can be defined as the composition of: i) a structural part  
 135 which directly models the predictand expected dependence from the predictors, and ii) an additive or multiplicative random  
 136 error  $\mathcal{E}$ . From the probability distribution of this random error, it is possible to derive the conditional probability distribution  
 137 of the predictand. In fact, both distributions are the same when an additive error model is considered, namely  
 138  $p(y - y_s, y_s) = p(e, y_s)$ . When this is the case, we rely on the following relation:

$$139 \quad y = y_s \left( \{ \theta_h, \theta_e \}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0 \right) + e \quad (1)$$

140 Equation (1) states that a predictand value  $y$  is obtained through the sum of a model prediction  $y_s$ , and the error, “residual”  
 141 or deviation  $e$ , which aggregates all sources of uncertainty. Model prediction,  $y_s$ , is expressed as a function of an observed  
 142 set of “k” predictors (or model inputs)  $\tilde{\mathbf{X}} = \{ \tilde{\mathbf{x}}_{1:N}^1, \dots, \tilde{\mathbf{x}}_{1:N}^k \}$ , the set of hydrological and error model parameters  $\{ \theta_h, \theta_e \}$ ,  
 143 and the initial conditions  $\tilde{\mathbf{s}}_0$ .

144 The aggregated error  $e$  can be decomposed in two components:

$$145 \quad e = \mu_{e|y_s} + \mathcal{E} \quad (2)$$

146 The first one in Eq. (2) is a systematic component,  $\mu_{e|y_s}$ , generally modeled as deterministic, and which could be non-  
 147 constant. The second component  $\mathcal{E}$ , is a random variable with zero mean and whose variance could also be variable. It is  
 148 very important to note that  $\mu_{e|y_s}$  is an error shifting function, which is only able to represent the expectations of the error  
 149 conditional distributions (the error conditional bias), namely  $\mu_{e|y_s} = E[e|y_s]$ , when the inferred errors fulfill the Total  
 150 Expectation Law (TEL), also called the Iterated Expectations Law, or Adam’s Law:

$$151 \quad E[e] = E \left\{ E[e|y_s] \right\} \quad (3)$$

152 where,  $E[\bullet]$  is the expectation operator. In other words, Eq. (3) tells us that the marginal (total) expectation of the error is  
 153 equal to the expectation of all error conditional (on  $y_s$ ) expectations.

154 Besides, Eq. (1) could also be written as:

$$155 \quad y = E[y|y_s] + \mathcal{E} \quad (4)$$

156 where  $E[y|y_s] = y_s + \mu_{e|y_s}$  is the deterministic part of the predictand, and  $\mathcal{E}$  is an additive random error. In the case that  
 157 a hydrological model would yield unbiased outputs then  $\mu_{e|y_s} = 0$  and so  $E[y|y_s] = y_s$ . It is important to note that we  
 158 can have a zero-mean error marginal distribution, but this is not equivalent to have hydrological model without bias, since  
 159 conditional biases  $\mu_{e|y_s} \neq 0$  could auto-compensate yielding a total expectation  $E[e] = 0$ . Bias is a non-random deviation  
 160 in the simulated  $y_s$  value, due mainly to hydrological model structural deficiencies and systematic errors in forcing input  
 161 data.



162 In this research, and without loss of generality, the basin outlet discharge is considered as the simulated variable of predictive  
 163 interest. In fact, this is the general practical situation and we have selected it as the only explicative variable for the bias. In  
 164 particular, we have assumed a straightforward double linear bias model defined by:

$$\begin{aligned}
 \mu_{e|y_s} &= \gamma && \text{if } y_s \leq y_0 \\
 \mu_{e|y_s} &= \gamma + \tau(y_s - y_0) && \text{if } y_s > y_0
 \end{aligned} \tag{5}$$

166 where  $\gamma$ ,  $\tau$  and  $y_0$  are error model parameters to be inferred jointly with the hydrological ones (in fact, all error model  
 167 parameters will be jointly inferred with the hydrological ones). This double function aims at the consideration of two  
 168 different expected error behaviors distinguishing the low flows (with a constant bias) from the high flows (with a linear bias).

169 As aforementioned, errors could exhibit a non-constant variance. Following Schoups and Vrugt (2010), Evin et al. (2013,  
 170 2014) and others, we assume in this paper a linear heteroscedasticity model, where  $y_s$  is the only explicative variable for the  
 171 conditional error standard deviation  $\sigma_{e|y_s}$ . In this case, we can write:

$$\sigma_{e|y_s} = \alpha + \kappa y_s \tag{6}$$

173 where  $\alpha$  and  $\kappa$  are error model parameters. It is important to note that function  $\sigma_{e|y_s}$  represents to the standard deviation  
 174 of the error conditional distributions, namely  $\sigma_{e|y_s}^2 = V[e|y_s]$ , only if the inferred errors fulfill the Total Variance Law  
 175 (TVL), also called the Variance Decomposition Law or Eve's Law:

$$V[e] = E\{V[e|y_s]\} + V\{E[e|y_s]\} \tag{7}$$

177 where  $V[\cdot]$  is the variance operator. Equation (7) tells us that the marginal (total) variance of the error is equal to the sum of  
 178 two terms: the first term assesses the expectation of the conditional (on  $y_s$ ) variances of the error, and the second term  
 179 evaluates the variance of the error conditional (on  $y_s$ ) biases. On the contrary, when TVL is not fulfilled, function  $\sigma_{e|y_s}$   
 180 does not represent to the standard deviation of the error conditional distributions, and an error standardization process by  
 181 using this function would be a simple error scaling, instead of a correct standardization.

182 Having defined the models for the error bias and the error variance, we can also consider the possibility that the random  
 183 component of the errors,  $\varepsilon = e - \mu_{e|y_s}$ , still exhibits serial correlation. Theoretically, the more accurate the bias model is,  
 184 the smaller the remaining error serial correlation shall be. This dependence (error autocorrelation) can be due to the  
 185 “memory” effect, caused by the propagation of forcing and structural errors through model storage components (Kavetski et  
 186 al., 2003). The error autocorrelation can be modeled, as in Schoups and Vrugt (2010), Evin et al. (2013, 2014) and others,  
 187 using an autoregressive (AR) model. At this stage the modification proposed by Evin et al. (2013) about the methodology  
 188 applied in Schoups and Vrugt (2010) is considered: that is, errors should be studentized before applying an autoregressive  
 189 error model on them. According to Evin et al. (2013), the reason for this change in the original method lies in the  
 190 mathematical behavior of the autoregressive equations, and is particularly related to the error accumulation properties which  
 191 are different in both approaches (for more details see Evin et al. (2013)). So, the standardized errors are defined as:

$$\eta = \sigma_{e|y_s}^{-1} (e - \mu_{e|y_s}) = \sigma_{e|y_s}^{-1} \varepsilon \tag{8}$$



193 where  $\mathcal{E}$  is the zero-mean additive random error,  $\sigma_{e|y_s}$  is the modeled standard deviation of the error conditioned on the  $y_s$   
 194 value and  $\mu_{e|y_s}$  is the modeled mean of the error (actually its bias) conditioned also on  $y_s$ .

195 Coming again to the matter of the errors autocorrelation, the equations of pure autoregressive model have the following  
 196 compact form (Box and Jenkins, 1994):

$$197 \quad \left(1 - \sum_{i=1}^p \phi_i B^i\right) \eta = \phi_p(B) \eta = z \quad (9)$$

198 which expresses the  $p$ -order autoregressive model over the standardized errors  $\eta$ , where  $\phi_i$  are the  $i$ -order autoregressive  
 199 coefficients,  $B$  is the backshift operator  $B^i \eta_t = \eta_{t-i}$  and  $z$  are the resulting innovations. Innovations are mutually  
 200 independent random errors which represent the measurement errors (e.g. in forcing data and discharge measures). They  
 201 follow a probability distribution with a constant variance  $\sigma_z^2$ . This variance, which should be considered as another  
 202 (derived) parameter of the error model, can be evaluated according to the following expression (Box and Jenkins, 1994):

$$203 \quad \sigma_z^2 = \sigma_\eta^2 \left(1 - \sum_{i=1}^p \rho_i \phi_i\right) \quad (10)$$

204 where  $\rho_i$  are the  $i$ -lagged autocorrelation coefficients of the standardized error series and  $\sigma_\eta^2$  is the variance of the  
 205 standardized errors. In this research we have used an AR(1) model as in Schoups and Vrugt (2010) and Evin et al. (2013,  
 206 2014). Since the innovations  $z$  are not standardized, they are subjected to a final transformation using their standard  
 207 deviation:

$$208 \quad a = \sigma_z^{-1} z \quad (11)$$

209 where  $\sigma_z$  is obtained from Eq. (10) and  $a$  are the final standardized innovations, which is an independent random variable  
 210 with zero mean and unit variance (namely a standard white noise). Substituting Eq. (8) in Eq. (9) and this in Eq. (11) we  
 211 obtain the relation between errors and innovations:

$$212 \quad a = \sigma_z^{-1} \phi_p(B) \left( \sigma_{e|y_s}^{-1} (e - \mu_{e|y_s}) \right) \quad (12)$$

213 At this point, we should establish the pdf of the standardized innovations obtained in Eq. (12). This research models the  
 214 standardized innovations using the Skew Exponential Power SEP(0,1,  $\xi$ ,  $\beta$ ) pdf, with two parameters (skewness  $\xi$  and  
 215 kurtosis  $\beta$ ). The SEP(0,1,  $\xi$ ,  $\beta$ ) pdf, as shown in Schoups and Vrugt (2010), offers great flexibility by avoiding the a  
 216 priori assumptions about specific forms of the innovations probability distribution. In fact, a SEP(0,1,  $\xi$ ,  $\beta$ ) pdf may adopt  
 217 a variety of forms, from the Normal to Laplace distribution, as well as it is capable to reproduce asymmetries and heavy tails.  
 218 The analytical expression of the SEP(0,1,  $\xi$ ,  $\beta$ ) pdf is defined as follows:

$$219 \quad p(a | \xi, \beta) = \frac{2\sigma_\xi}{\xi + \xi^{-1}} w_\beta \exp\left(-c_\beta |a_\xi|^{\frac{2}{1+\beta}}\right) \quad (13)$$

220 where



$$221 \quad a_{\xi} = \xi^{-\text{sign}(\mu_{\xi} + \sigma_{\xi} a)} (\mu_{\xi} + \sigma_{\xi} a) \quad (14)$$

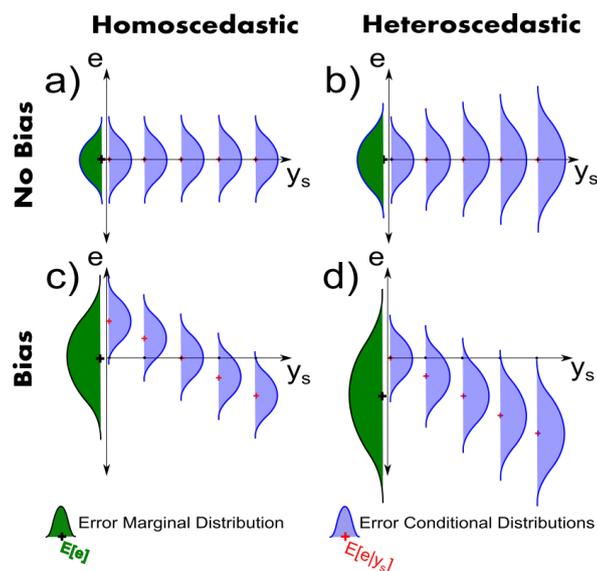
222 is a function of the standardized innovations  $a$ , and  $\mu_{\xi}, \sigma_{\xi}, c_{\beta}, w_{\beta}$  are a function of skewness  $\xi$  and kurtosis  $\beta$   
 223 parameters, as shown in Schoups and Vrugt (2010).

224 As summary, the vector of error model parameters is  $\theta_e = \{\alpha, \kappa, \gamma, \tau, y_0, \phi_1, \xi, \beta\}$  and  $\sigma_z$  could be considered as a  
 225 derived parameter. A major question related with the error model parameters must be pointed out. The application of the TLs  
 226 on the error bias and variance models reduces the degrees of freedom in the Bayesian inference problem, and consequently,  
 227 also reduces the number of independent error model parameters, namely the dimension of the parameter space. This will be  
 228 clearly exposed in section 4.4.

### 229 **2.3 Why and when is imperative the enforcement of the Total Laws**

230 Firstly, we want to emphasize a key concept which was previously exposed in the introduction: predictive uncertainty must  
 231 be linked with the existence of a joint probability distribution of the predictand observations and the related model  
 232 predictions. Under the hypothesis of an additive error model, this will be equivalent to considering the joint probability  
 233 distribution of the predictand errors and the model predictions, as it was made for example in Montanari and Brath (2004),  
 234 albeit they worked the statistics in the NQT (Kelly and Krzysztofowicz, 1997) transformed space. We are actually modeling  
 235 the conditional distribution of the error, given the model prediction. In other words, we are modeling indirectly the above  
 236 mentioned joint probability distribution of errors and model predictions. This means that the error conditional distributions  
 237 must fulfill the proper restrictions, in order to ensure that all of them make up the full joint distribution. The enforcement of  
 238 the Total Laws (TLs) exposed in Eqs. (3) and (7) allows us to impose the necessary restrictions in order to achieve that the  
 239 conditional and marginal distributions of the inferred errors belong to the same bidimensional joint distribution of these  
 240 errors and the modeled state variable of interest (the simulated streamflow in this research).

241 An incorrect error model will yield biased hydrological parameters (Sorooshian and Gupta, 1983) and, obviously, an  
 242 incorrect uncertainty assessment. Two main causes can lead to this situation: i) an incorrect hypotheses about the error  
 243 conditional distribution features (shape, mean and variance) or about the errors dependence structure; ii) making inferences  
 244 without enforcing the TLs when we are modeling the error and state variable joint distribution through the definition of its  
 245 conditional distributions. In order to better understand the TLs importance and implications, Fig. 1 shows what the TLs entail  
 246 under different hypotheses for the error variance and bias models.



247

248 **Figure 1. Schemes of decomposition of the state variable ( $y_s$ ) and its error joint distribution, in its conditional (blue) and marginal**  
 249 **(green) distributions, under different bias and variance error models.**

250 Fig. 1-(a) shows a scheme which is representative of the SLS error model, with the hypotheses of homoscedasticity (errors  
 251 with constant variance) and also a constant zero-bias. Therefore, all conditional distributions are a “copy” of the error  
 252 marginal distribution. Both hypotheses implicitly entail  $E[e] = E[e|y_s] = 0$  and  $V[e] = V[e|y_s]$ ; i.e. the TLs given  
 253 by Eqs. (3) and (7) are fulfilled, without the necessity of enforcing them.

254 Fig. 1-(b) represents the Weighted Least Squares error model (WLS), with non-constant error variance and constant zero-bias  
 255 hypotheses. This scheme shows the conditional distributions assuming a linear relationship between the errors standard  
 256 deviation and the simulated state variable (similar to that proposed in Eq. (6)). With WLS it is needed to explicitly enforce  
 257 the TVL on the variance model, in order to get an error marginal distribution with a variance equal to the average variance of  
 258 the error conditional distributions. For this particular error model (WLS), Eq. (7) collapses in the expression

$$259 \quad V[e] = E\{V[e|y_s]\}.$$

260 Fig. 1-(c) shows the case of having errors with a constant conditional variance and a non-constant linear bias. In this case, the  
 261 marginal distribution is different from the conditional ones, although all these have the same shape. This is due to the non-  
 262 constant bias which affects the position of the conditional distributions. Moreover, it can occur (as in the example in Fig. 1-  
 263 (c)) the effect of errors autocompensation: the marginal distribution shows a zero-mean value whereas the conditional  
 264 distributions exhibit bias. In this case, a non-constant conditional bias means that Eq. (3) for the TEL is not automatically  
 265 fulfilled (must be enforced). With a constant conditional variance, the TVL in Eq. (7) simplifies to  
 266  $V[e] = V[e|y_s] + V\{E[e|y_s]\}.$

267 Finally, Fig. 1-(d) shows the general case with both non-constant error conditional variances and bias. In this case, we must  
 268 enforce on both variance and bias models the general expressions of the TLs (Eqs. (3) and (7)). Section 4 describes the  
 269 inference settings for the case study and how the TLs are applied on each of them. On the contrary, section 6.1 will show the  
 270 consequences of not enforcing the TLs.



271 **3. Inference**

272 **3.1 Parameter uncertainty**

273 In Bayesian inference, it is necessary to construct a probability model based on the observed data of the variable of interest,  
 274 as well as on the model parameters. Formal inference requires the setting up of a joint probability model to take all these  
 275 random variables (i.e. observed state variables and model parameters) into account. Therefore, given the hydrological and  
 276 error models,  $\mathfrak{M}_{h,e}$ , and also given the set of initial conditions,  $\tilde{\mathbf{s}}_0$ , and an observation of the “k” predictors,  
 277  $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_{1:N}^1, \dots, \tilde{\mathbf{x}}_{1:N}^k\}$ , it must be constructed  $p(\tilde{\mathbf{y}}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$ , the joint distribution of observations of the state variable of  
 278 interest,  $\tilde{\mathbf{y}}$ , and the hydrological and error model parameters,  $\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}$ . This joint distribution, also called “the full  
 279 Bayesian model” (Bolstad, 2010), is conditioned on  $\mathfrak{M}_{h,e}$ ,  $\tilde{\mathbf{s}}_0$ ,  $\tilde{\mathbf{X}}$  and is formed by taking the product of two probability  
 280 distributions:

$$281 \quad p(\tilde{\mathbf{y}}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = p(\tilde{\mathbf{y}} | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \tilde{\mathbf{y}}) p(\tilde{\mathbf{y}}) \quad (15)$$

282 Then, based on Eq. (1) that hypothesizes an additive error model, we have  $p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \tilde{\mathbf{y}}) = p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \mathcal{E})$ , where  $\mathcal{E}$  is  
 283 the random error solved from Eq. (2). The analytical expression for  $p(\mathcal{E} | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$  is called the sampling distribution,  
 284 when the parameters are considered as known. The same analytical expression is called the parameter likelihood function  
 285  $\ell(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \mathcal{E})$ , when the known variables are the observed data. The likelihood function incorporates the effect of the data  
 286 on the prediction model, namely on the hydrological and error models jointly considered. Besides, the parameter prior  
 287 distribution  $p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$  describes the a priori knowledge we have about the hydrological and error models. Then, by  
 288 applying Bayes’ Rule in its unscaled form on Eq. (15), the parameter posterior distribution is given by:

$$289 \quad p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \mathcal{E}) \propto \ell(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \mathcal{E}) p(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (16)$$

290 Since in section 2.2 we have already modeled the errors, we can derive from it the analytical expression for the likelihood  
 291 function. This task is detailed in Appendix A, where the following log-likelihood function is obtained:

$$292 \quad \mathcal{L}(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \mathcal{E}) \cong N \log \frac{2\sigma_\xi w_\beta}{\sigma_z (\xi + \xi^{-1})} - \sum_{t=1}^N \log \sigma_{\varepsilon|y_t} - c_\beta \sum_{t=1}^N |a_\xi|^{1+\beta} \quad (17)$$

293 where:  $\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}$  is the set of hydrological and error model parameters;  $\mathcal{E}$  is the random error;  $\sigma_{\varepsilon|y_t}$  is the conditional  
 294 standard deviation of these random errors, which can be estimated by Eq. (6);  $\sigma_z$  is the standard deviation of innovations,  
 295 estimated with Eq. (10); the terms  $\mu_\xi, \sigma_\xi, c_\beta, w_\beta$  are a function of the skewness  $\xi$  and kurtosis  $\beta$ , parameters of the SEP  
 296 distribution (more details in Schoups and Vrugt (2010));  $a_\xi$  is a function of the standardized innovations given by Eq. (14);  
 297 and  $N$  is the number of observations.

298 Parameter posterior pdf describes the probability density that the model parameters assume particular values conditioned on  
 299 the observed data. I.e. it describes the parameter uncertainty. In this study, both hydrological and error model parameters are  
 300 conditioned simultaneously (joint inference) on observations. Computational sampling of the posterior is made through a  
 301 Markov chain Monte Carlo (MCMC) algorithm. Modern MCMC algorithms, which are an evolution of the original



302 Metropolis algorithm (Metropolis and Rosenbluth, 1953), are currently used to “calibrate” the model parameters and to assess  
 303 the uncertainties of these parameter estimates. Even high-dimensional and multimodal posterior distributions can be  
 304 efficiently sampled by these recent MCMC algorithms after convergence of the chain (or chains) has been reached (Cowles  
 305 and Carlin, 1996). The Differential Evolution Adaptive Metropolis algorithm DREAM-ZS (Schoups and Vrugt, 2010; Laloy  
 306 and Vrugt, 2012) has been selected in this research to sample the posterior distribution of the parameters. The DREAM-ZS  
 307 algorithm is a modification of the DREAM (Vrugt et al., 2008, 2009a) algorithm. DREAM-ZS simultaneously runs multiple  
 308 Markov chains to explore the whole parameter space, uses an archive of past states to generate candidate points in each  
 309 individual chain and automatically adjusts the scale and orientation of the proposal distribution. Such sampling is more  
 310 efficient than an optimal random walk Metropolis. The R-statistic has been used to check whether the chains have converged,  
 311 with a threshold value of R equal to 1.2 (Gelman and Rubin, 1992). In this research, the number of chains used in the  
 312 inference configurations was 10.

### 313 3.2 Predictive Uncertainty

314 One of the main advantages to pursuing a formal method of Bayesian inference is the ability to make a correct estimate of the  
 315 predictive uncertainty. Given the hydrological and error models,  $\mathfrak{M}_{h,e}$ , a formal definition of the predictive uncertainty for the  
 316 prediction in a period with  $N$  time steps, adapted from Mantovan and Todini (2006) is given by the next equation:

$$317 \quad p\left(y \mid y_s(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0)\right) = \int p\left(y \mid y_s(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}, \tilde{\mathbf{X}})\right) p\left(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} \mid \tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0\right) d\boldsymbol{\theta} \quad (18)$$

318 where the left term  $p\left(y \mid y_s(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0)\right)$  is the probability density of the predictand  $y$ , conditioned upon its  $N$   
 319 observations  $\tilde{\mathbf{y}}$ , the set of initial conditions  $\tilde{\mathbf{s}}_0$ , the observed predictors  $\tilde{\mathbf{X}}$ , and necessarily also on  $\mathfrak{M}_{h,e}$ . Hence, the  
 320 dependence on the parameters has been marginalized integrating over the entire domain of existence of these parameters; i.e.,  
 321 the ensemble of all possible parameter realizations. In other words, for assessing the predictive distribution (hereinafter PD)  
 322 of the predictand, one has to take into account all possible model predictions (namely one per parameter vector realization),  
 323 instead of only considering the prediction which corresponds to the most plausible parameter set. The first right term  
 324  $p\left(y \mid y_s(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}, \tilde{\mathbf{X}})\right)$  is the probability density of the predictand conditioned on the observed predictors and a unique  
 325 set of parameters. The last right term  $p\left(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} \mid \tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0\right)$  is the posterior of the parameters given the observed  
 326 predictors, the observed predictand and the initial conditions. This probability distribution can be derived by means of a  
 327 Bayesian inferential process (as explained in section 3.1), and it is used to marginalize out the conditionality on the  
 328 parameters of the prediction.

329 The way of computing the PD is as follows. First a unique set  $\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}$ , of hydrological and error model parameters, is  
 330 sampled from their posterior distribution. After the parameter posterior sampling, the hydrological model is run for the  
 331 sampled hydrological parameters  $\boldsymbol{\theta}_h$ , getting a model simulation of the predictand  $y_s(\boldsymbol{\theta}_h, \tilde{\mathbf{X}})$ . Then, by using repeatedly  
 332 the prediction sampling equation given by:

$$333 \quad \hat{y} = y_s\left(\boldsymbol{\theta}_h, \tilde{\mathbf{X}}\right) + \mu_{e|y_s} + \sigma_{e|y_s} \sigma_z \phi_p^{-1}(B)a \quad (19)$$

334 we get the probability density of the predictand, conditioned upon a unique set of hydrological and error model parameters,  
 335 namely  $p\left(y \mid y_s(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}, \tilde{\mathbf{X}})\right)$ , where  $\hat{y}$  is an estimated prediction of the predictand, conditioned on the previously



336 obtained model simulation  $y_s(\boldsymbol{\theta}_h, \tilde{\mathbf{X}})$  (i.e. on the hydrological model) and also conditioned on a unique realization from  
 337 the error model. This error model realization is obtained through the previously sampled error model parameters  $\boldsymbol{\theta}_e$ , using a  
 338 random extraction  $\mathbf{a}$  from the SEP innovations distribution. To define with enough resolution this pdf of the predictand  
 339 conditioned on  $y_s(\boldsymbol{\theta}_h, \tilde{\mathbf{X}})$ , a sufficient number “s” of prediction samples  $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_{1:N}^1, \dots, \hat{\mathbf{y}}_{1:N}^s\}$  must be obtained with the  
 340 corresponding “s” random extractions from the SEP innovations distribution. Finally, all previous steps are repeated over  
 341 multiple hydrological and error parameter posterior samples  $\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}_{i=1, \dots, m}$ , generating an ensemble of predictions  
 342  $\{\hat{\mathbf{Y}}\}_{i=1, \dots, m}$ , in which the dependence on the parameters has been marginalized. The present paper considers the mean of that  
 343 ensemble of predictions as the estimator for the expected prediction  $\hat{\mathbf{y}}^*$ , unconditioned to the parameters, formally:

$$344 \quad E\left[y \mid y_s(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{s}}_0)\right] \approx \hat{\mathbf{y}}^* = \frac{1}{ms} \sum_{i=1}^m \sum_{j=1}^s \hat{\mathbf{y}}_{1:N}^{i,j} \quad (20)$$

#### 345 4. Application to Rainfall-Runoff modeling

346 This paper considers the simulated daily streamflow  $q_s$ , as the state variable of interest to be predicted. The key concept is  
 347 that, this state variable and its error belong to a bivariate population with a unique joint pdf. This belonging relationship  
 348 imposes restrictions which must be fulfilled by both the state variable and the errors; actually by the parameters of  
 349 hydrological and error models. That is to say, the error and hydrological inferred parameters, which generally are only  
 350 conditioned by the observed data through the likelihood function, will also be conditioned by this belonging relationship  
 351 expressed by the TLs in Eqs. (3) and (7). So, this section shows how to correctly perform the parameter joint inference of  
 352 hydrological and error models, in order to reach the fulfillment of the TLs.

##### 353 4.1 The basin

354 We have selected the French Broad River (hereinafter FB) basin. It is located in North Carolina (USA) and has an area of  
 355 2448 km<sup>2</sup>. FB basin information can be obtained from the Model Parameter Estimation Experiment (MOPEX) data set (Duan  
 356 et al., 2006). Besides, FB basin is a broadly known basin by the hydrologists’ community. Previous research related with this  
 357 paper also performed inferences in this basin, as the works of Schoups and Vrugt (2010) or Evin et al. (2013, 2014). This  
 358 basin is a representative humid catchment where actual evapotranspiration is energy-limited, hence runoff production  
 359 opportunity is high: FB basin has a mean annual rainfall of 1495 mm, and a mean annual potential evapotranspiration of 820  
 360 mm. Five years (from 01/01/1962 to 31/12/1966) of observed daily forcings (precipitation and potential evapotranspiration)  
 361 and observed daily streamflow were used to identify hydrologic and error model parameters. The two previous years were  
 362 used as warm-up period.

##### 363 4.2 The hydrological models

364 Two hydrological models have been chosen to illustrate the theoretical issues on which this research deals with. The reason  
 365 for the election is mainly their use in researches related with the present paper.

366 The first one is the CRR model, which was also used in Schoups and Vrugt (2010) or Schoups et al. (2010). CRR is a  
 367 conceptual spatially lumped hydrologic model for the simulation of daily rainfall-runoff processes, and it has seven  
 368 parameters to calibrate: the maximum interception ( $I_{\max}$ , units of mm), the Soil water storage capacity ( $S_{\max}$ , units of mm),  
 369 the maximum percolation rate ( $Q_{\max}$ , units of mm/d), the evaporation parameter ( $\alpha_e$ , dimensionless), the runoff parameter  
 370 ( $\alpha_r$ , dimensionless), the fast reservoir constant ( $K_f$ , units of days) and the slow reservoir constant ( $K_s$ , units of days). The  
 371 model considers the processes of interception, throughfall, evapotranspiration, runoff generation, percolation, and surface and  
 372 subsurface routing of water to the basin outlet.



373 The second model is the well-known GR4J. It is also spatially lumped and was developed to provide, on average, good  
374 performance across a wide range of catchment conditions (Perrin et al., 2003). The GR4J model, with only four parameters,  
375 is a more parsimonious model than CRR. These four parameters are: the production store capacity ( $\theta_1$ , units of mm), the  
376 groundwater exchange coefficient ( $\theta_2$ , units of mm), the one day-ahead maximum capacity of the routing store ( $\theta_3$ , units of  
377 mm) and the base time of the unit hydrograph ( $\theta_4$ , units of days).

378 In order to infer the parameter posteriors (as explained in section 3.1), we have assumed non-informative Priors. The  
379 assumption of uniform Priors for all parameters is acceptable since a sufficient number of daily data points ( $N=1825$ )  
380 supports the parameter inference.

381 An important question to be remarked is that snow accumulation and snowmelt are not accounted for in any of both models,  
382 although these processes occur in the FB basin, but with low significance. Therefore, it is expected that this model  
383 misspecification will have a non-severe influence in the results, but also an undermining of either the reliability of the  
384 inferred parameters or the reliability of the prediction. These consequences will be evaluated in the performed inferences.

### 385 4.3 Implemented error models

386 In this paper, we are going to perform different inferences considering four error models which, sorted in crescent order of  
387 complexity, are: SLS, WLS, GL++ and GL++Bias. Besides, for the WLS and GL++ error models, the inferences will be also  
388 performed with and without the enforcement of the TLs. This later inferences are called NTL inferences and they are an  
389 illustrative example of the problems which can arise from the joint inference when TLs are neglected.

390 The SLS (Standard Least Squares) error model is the standard approach in regression analysis. For SLS, the errors are a zero-  
391 mean measurement random noise, so they are serially independent and, normally and identically distributed. SLS is  
392 considered as the reference inference, since it is the most common method in parameter estimation.

393 The WLS approach (Weighted Least Squares) considers all the same hypotheses that SLS, except that WLS applies a  
394 heteroscedastic error model which weights the errors by their standard deviation as in Eq. (6). WLS needs to apply the TVL.  
395 However, section 6.1 will show what happens with the WLS inference without the enforcement of the TVL (which is called  
396 WLS-NTL inference).

397 The GL++ is an evolution of the GL error model described in Schoups and Vrugt (2010). The modifications on GL are two.  
398 Firstly we have included the recommendation given by Evin et al. (2013): errors must be studentized before applying an  
399 autoregressive model on them. Secondly, the joint inference has taken into account the enforcement of the TLs. Therefore,  
400 GL++ assumes zero-mean errors and the same heteroscedasticity model as in the WLS, but relaxes the hypothesis of error  
401 independence through a first order AR model (Eq. (9)). The resulting innovation has a variance given by Eq. (10) and is  
402 distributed according to a SEP distribution described by Eqs. (11) to (14). The Log-likelihood function in Eq. (17) includes  
403 all these assumptions, assuming an unbiased conditional error. Again, section 6.1 will show the results of the GL++NTL  
404 inference, in order to illustrate what occurs to this kind of error model when TVL is not enforced.

405 Finally, the GL++Bias error model assumes all hypotheses of GL++, but relaxes the hypothesis of having zero-mean  
406 conditional error distributions. A bias model is hypothesized according to Eq. (5), which tries to correct these conditional  
407 biases. Therefore, GL++Bias is the full error model described in section 2.2.

### 408 4.4 Enforcement of the Total Laws

409 On all the above described error models, it will be applied the enforcement of either one or both TLs, except on SLS (because  
410 it is not needed) and on the NTL cases. In the case of WLS and GL++ approaches, they must yield inferred errors which  
411 fulfill the simplified expression of the TVL given by:

$$412 \quad V[e] = E\left\{V[e|q_s]\right\} \quad (21)$$



413 This conditioning to TVL allows us to reduce one degree of freedom in the inference problem: the intercept  $\alpha$  from the  
 414 heteroscedasticity linear model will be a function of other parameters and the state variable. So, this error parameter does not  
 415 need to be sampled in the computational Bayesian inference process. Following the calculations in Appendix B, we obtain:

$$416 \quad \alpha = \left( V[e] - \kappa^2 V[q_s] \right)^{0.5} - \kappa E[q_s] \quad (22)$$

417 Eq. (22) is the condition that must meet the intercept parameter  $\alpha$  in order to achieve that the inferred errors with the WLS  
 418 and GL++ approaches, are able to fulfill the TVL.

419 In the case of the GL++Bias error model, both error variances and bias are supposed to be non-constant. In this case we must  
 420 enforce on both variance and bias models the general expressions of the TLs given by Eqs. (3) and (7). In this way, this  
 421 enforcement (as shown in Appendix B) allows us to set the following four parameters:

422 For the linear heteroscedasticity model given by Eq. (6),

423 i) The intercept parameter,  $\alpha$

$$424 \quad \alpha = \left( V[e_1] - \kappa^2 V[q_{s_1}] \right)^{0.5} - \kappa E[q_{s_1}] \quad (23)$$

425 ii) The slope parameter,  $\kappa$

$$426 \quad \kappa = \left( \frac{V[e_1] - V[e_2] + \tau^2 V[q_{s_2}]}{E[q_{s_1}^2] - E[q_{s_2}^2]} \right)^{0.5} \quad \text{for } \alpha \approx 0 \quad (24)$$

427 It is important to note that the later equation, which yields  $\kappa$  explicitly, is only valid for small values of parameter  $\alpha$ . If  
 428 not, parameter  $\kappa$  should be calculated in an iterative way by solving Eq. (B13).

429 And for the double linear bias model given by Eq. (5),

430 iii) The intercept parameter,  $\gamma$ , for the low-flows population

$$431 \quad \gamma = E[e_1] \quad (25)$$

432 iv) The slope parameter,  $\tau$ , for the high-flows population

$$433 \quad \tau = \frac{E[e_1] - E[e_2]}{q_0 - E[q_{s_2}]} \quad (26)$$

434 where the subindexes 1 and 2 distinguish the low and high streamflow populations respectively. The four Eqs. (23) to (26)  
 435 allow that inferred errors by GL++Bias be able of fulfilling both TLs. Additionally, these imposed equations remove four  
 436 degrees of freedom from the inference problem. That is to say, GL++Bias has only four (and not eight) free error model  
 437 parameters to be sampled by the MCMC algorithm. These free parameters are: the bias parameter  $q_0$ , the SEP parameters  $\beta$   
 438 and  $\xi$ , and the autoregressive coefficient  $\phi_1$ .

439



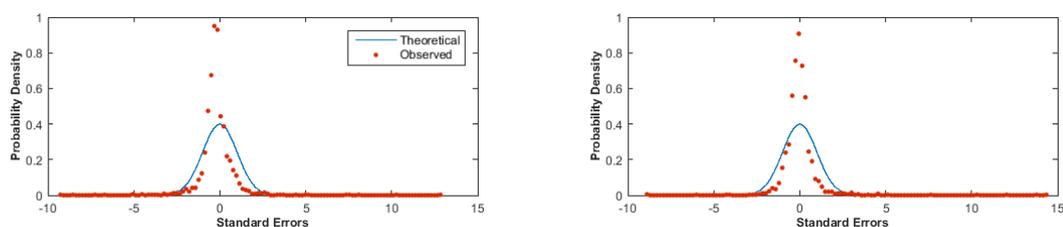
## 440 5. Results

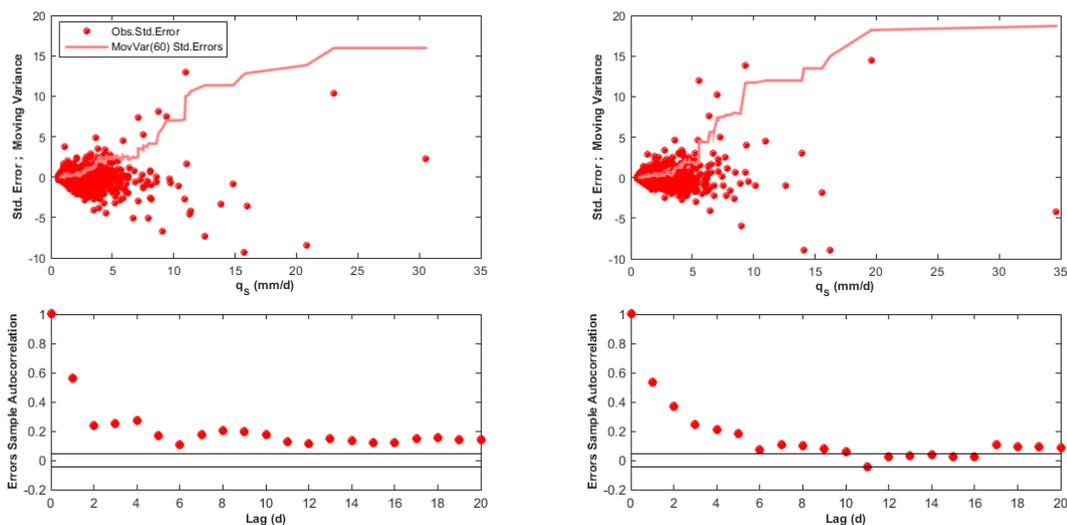
### 441 5.1 The reference inference: SLS

442 The target of the SLS method is the mean squared error minimization of the prediction, and errors are assumed to be  
443 independent, homoscedastic and Gaussian with zero-mean. Therefore, a SLS calibrated hydrological model must yield the  
444 best fitting to the observed values. With this aim, the SLS method tries to compensate the model misspecifications and the  
445 data errors through the parameter values, by a departure from their “true” value. We consider the parameter “true” values as  
446 the generally unknown parameter set, which would be inferred in a calibration without model structural deficiencies or biased  
447 data errors: i.e. only with random measurement errors.

448 SLS yields error marginal variances in the case study for both hydrological models ( $0.34 \text{ mm}^2/\text{d}^2$  and  $0.43 \text{ mm}^2/\text{d}^2$  for CRR  
449 and GR4J respectively) which are the minimum reached in all performed inferences (compare RMSE values in Table 1). This  
450 situation can be expected, since this is the way to reach the target of the SLS inference. Another interesting feature is the  
451 (marginal) mean of the errors, which is not zero (i.e., SLS unbiased errors hypothesis is not fulfilled), and it is slightly greater  
452 with CRR than with GR4J ( $-0.11 \text{ mm}/\text{d}$  and  $-0.02 \text{ mm}/\text{d}$  respectively). In fact, as it was also expected, SLS inference does not  
453 fulfill any of its underlying hypotheses about the errors (Fig. 2): normality, homoscedasticity and independence. Regarding  
454 the normality assumption (Fig. 2, top panels), the error distribution has an excess of kurtosis, for both hydrological models.  
455 The non-compliance of the homoscedasticity is clearly shown in middle panels of Fig. 2: the funnel-shaped scatter plot of the  
456 standardized errors versus the simulated streamflow shows that error variance varies with the streamflow magnitude. To  
457 assess this heteroscedastic property numerically, it has been calculated a moving variance of the standardized errors, by  
458 taking a moving window through the scatter plot, with a length of 60 data, over which the variance is calculated. The result of  
459 this calculation is the red line drawn over the scatter. The flatness of this line would indicate the homoscedasticity of the  
460 errors and the line is far to be flat. Finally, the errors independence hypothesis can be judged through the autocorrelation  
461 function shown in Fig. 2 (bottom panels). These graphs also show the 95% Anderson’s limits (Anderson, 1942) for null  
462 correlation hypothesis. As it can be observed, SLS yields a clear not null autocorrelation for the first ten lags, for both  
463 hydrological models.

464 Table 1 shows the Log-likelihood value for each of the inferred error models and it is clear that the worst values are for SLS  
465 ( $-1625$  with CRR and  $-1819$  with GR4J hydrological model). On the contrary, as shown in Tables 2 and 3, SLS generally (if  
466 the parameters are identifiable) yields the parameter estimates with the lowest uncertainty: namely, with the smallest sample  
467 coefficients of variation (CV). In this case study, GR4J model shows perfectly identifiable inferred parameters and with little  
468 uncertainty. However, CRR model shows two non-identifiable parameters,  $\alpha_e$  and  $\alpha_t$ , which are highly uncertain parameters  
469 in comparison with the other in SLS (see Table 2), showing CVs of 0.20 and 1.77 respectively.



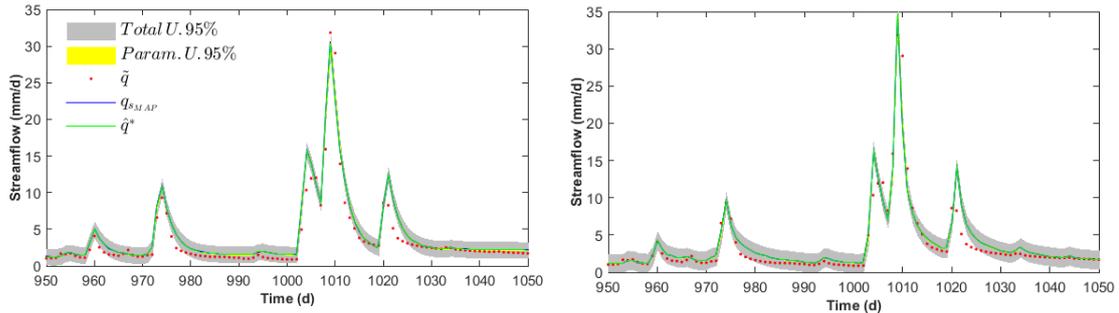


470 **Figure 2. Test of SLS error model hypotheses for CRR (left column) and GR4J (right column) hydrological models: normality of**  
 471 **observed standard errors in top panels, homoscedasticity evaluation of errors in middle panels and independence assessment of the**  
 472 **errors through the autocorrelation function in bottom panels (the 95% Anderson's limits in solid black lines).**

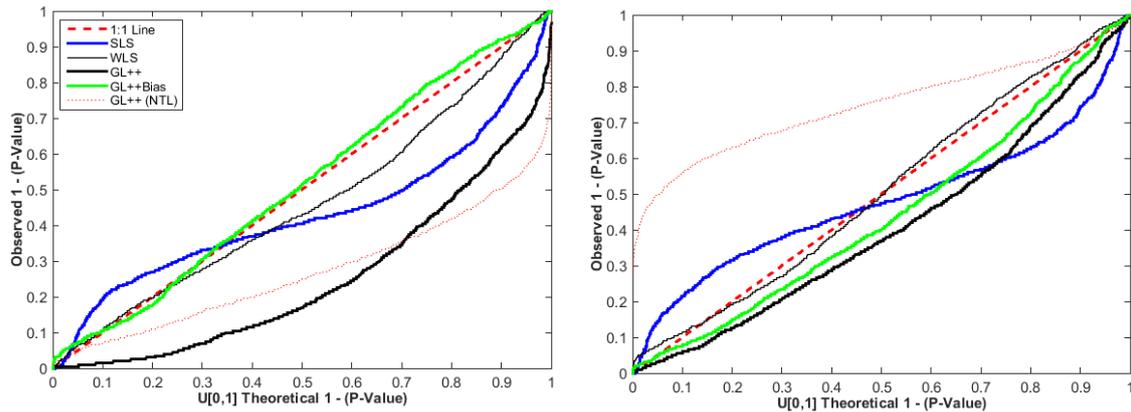
473 Concerning the performance of the mean prediction, Table 1 shows that SLS has the highest NSE index among all inferred  
 474 error models. On the one hand, it results a NSE=0.90 in CRR, whereas GR4J yields a NSE=0.87. On the other hand, CRR  
 475 shows a Volume Error, defined as  $VE = (V_s - V)V^{-1}$ , larger than the obtained with GR4J model. This result is coherent  
 476 with the value for the marginal mean of the errors obtained for each model. In Fig. 3 are represented the observations  $\tilde{q}$ , the  
 477 mean prediction  $\hat{q}^*$  and the hydrological model prediction for the Maximum a Posteriori (MAP) parameter set  $q_{s,MAP}$ . This  
 478 figure shows how CRR yields a nice fitting to the largest peakflow of the calibration period, but CRR obtains a moderate  
 479 fitting to the rest of peakflows and a general overestimation of the low flows. A similar general behavior is achieved by GR4J  
 480 model, as can be seen in the same Fig. 3.

481 The validation of the predictive uncertainty for both hydrological models can be visually checked in Fig. 3: this figure shows  
 482 the 95% prediction uncertainty band, resulting with a constant width for all the streamflow magnitudes (although a merely  
 483 visual effect seems to contradict this). It is important to note that Fig. 3 also shows the parameter uncertainty band, but its  
 484 extreme thinness means that, in this inference, the contribution of the parameter uncertainty to the total uncertainty can be  
 485 neglected.

486 Another more comprehensive method of testing the validity of the predictive uncertainty is through the analysis of the PD of  
 487 the predictand with the PP-Plots tool (Laio and Tamea, 2007; Thyer et al., 2009; Renard et al., 2010). In Fig. 4 (left panel) the  
 488 blue line shows a systematic bias (overprediction) with an uncertainty overestimation for the CRR model. In the same Fig. 4  
 489 (right panel), a very small overpredictive bias and also an uncertainty overestimation is shown for the GR4J model when SLS  
 490 inference is used.



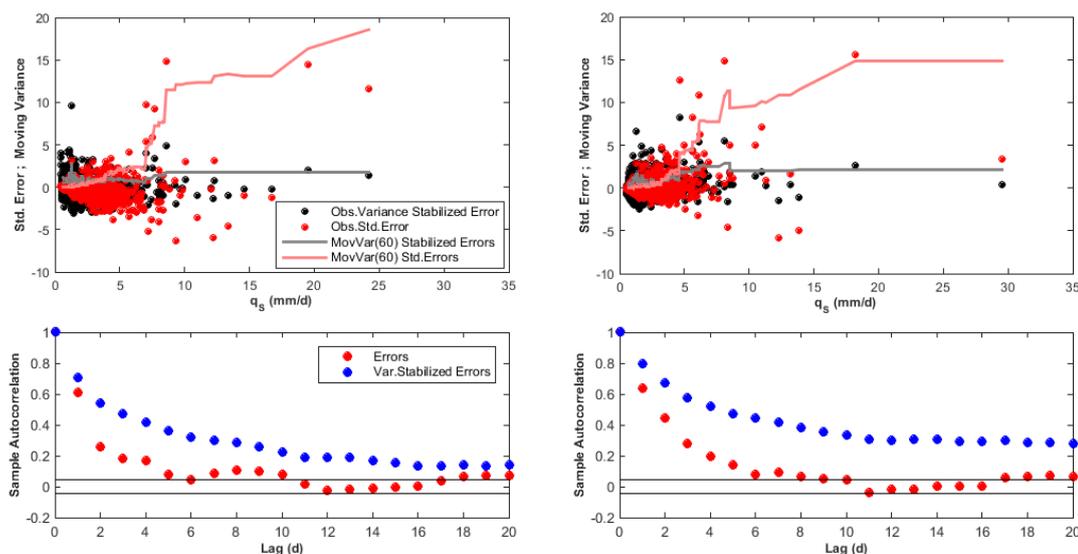
491 **Figure 3. Predictive performance and 95% uncertainty bands (parameter and total) with SLS inference on CRR (left) and GR4J**  
 492 **(right) models (parameter uncertainty band is hardly appreciable).**



493 **Figure 4. PP-Plots of the PDs for all performed inferences with CRR (left) and GR4J (right) models.**

494 **5.2 WLS inference**

495 In this case study, the WLS inference does not fulfill any of its error model hypotheses, as SLS, with the exception of the  
 496 homoscedasticity hypothesis. As in SLS inference, the errors are not Normal (Fig. not shown). The good variance  
 497 stabilization of the inferred errors with both hydrological models can be observed in top panels of Fig. 5, where are depicted  
 498 the scatter plots for the standard errors and for the stabilized errors. Over the scatters there are also two lines representing the  
 499 corresponding moving variances, calculated as it was explained for the SLS inference. As it can be observed, the calculated  
 500 moving variance for the errors has the same appearance as in SLS. But, the moving variance for the stabilized errors is  
 501 practically a horizontal line; i.e., it represents a constant variance. However, bottom panels in Fig. 5 show how the stabilized  
 502 errors have a higher autocorrelation than the raw errors. In other words, WLS solves the errors heteroscedasticity problem,  
 503 but increases the problem of error serial dependence. The fulfillment of the homoscedasticity hypothesis is able to improve  
 504 the Log-likelihood respect to SLS, in a 72% for CRR model whereas this improvement with GR4J is about 60% (Table 1).



505 **Figure 5.** Test of WLS error model hypotheses for CRR (left column) and GR4J (right column) models. In top panels,  
 506 homoscedasticity evaluation for the observed errors (red dots and red line) and for the variance stabilized errors (black dots and  
 507 grey line). In bottom panels, the independence assessment for observed errors and for the variance stabilized errors, through the  
 508 autocorrelation function (the 95% Anderson's limits in solid black lines).

509 The performance of the WLS mean prediction and the negligible contribution of the parameter uncertainty to the total  
 510 uncertainty are very similar to the SLS one (graphs not shown). Comparing both inferences in Table 1, WLS exhibits a  
 511 slightly worse NSE index, but a better VE. The reduction in NSE is due to the fact that WLS gives a minor weight to the high  
 512 flows during the calibration, yielding a poorer fit to the peakflows.

513 An improvement with respect to SLS is that WLS has a very good prediction uncertainty assessment, with GR4J slightly  
 514 better than with CRR, as it is shown in the PP-Plots of Fig. 4. But, the fact that WLS does not fulfill the error independence  
 515 hypothesis implies that: i) there is a problem with the hydrological model structure and/or with the data reliability; and ii) we  
 516 have a good uncertainty assessment of a distorted hydrological model which is being used with biased inferred parameters. It  
 517 is expected that, out of the calibration period, these biased parameters will induce a poor performance of the hydrological  
 518 model as well as of the uncertainty assessment. This was called model divergence phenomenon by Sage and Melsa (1971)  
 519 and Sorooshian and Dracup (1980), long time ago.

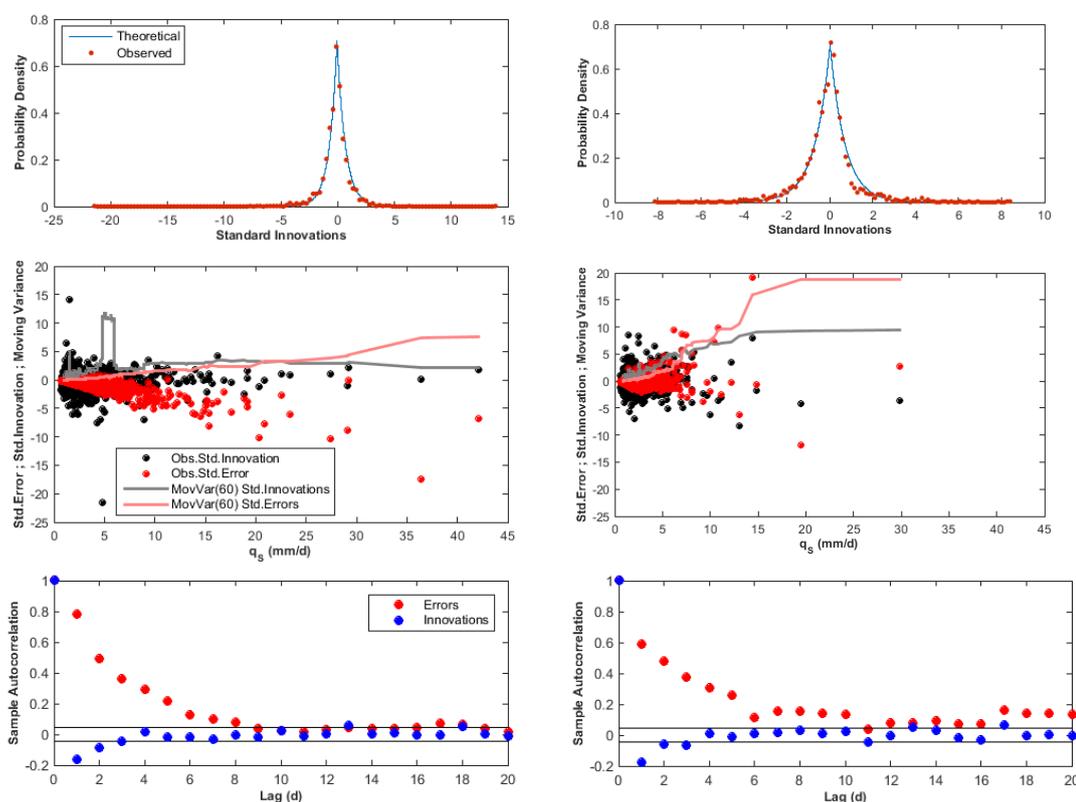
### 520 5.3 GL++ inference

521 The inference with the GL++ error model fulfills practically all its hypotheses with both CRR and GR4J hydrological  
 522 models. As it is shown in top panel of Fig. 6, CRR presents an error distribution with heavier tails than GR4J, but despite this  
 523 difference, the fitting to the SEP distribution is good and similar for both hydrological models. The error autocorrelation is  
 524 almost eliminated with both hydrological models, although as shown in Fig. 6 (bottom panel), still remains a very small but  
 525 significant negative autocorrelation of the innovations for the first three lag times.

526 Regarding the variance stabilization, it can be seen in Fig. 6 (middle panel) how the conditional variance of errors and  
 527 innovations (represented as previously explained by a moving sample variance) are very similar almost for all the range of  
 528 simulated streamflows. Hence, GL++ is not able to stabilize the error variances as WLS does it. The reason of this poor  
 529 behavior of the variance stabilization is the error bias which has arisen with this GL++ inference, which is not modeled. This  
 530 overpredictive bias can be clearly noticed for the CRR model: standard errors (red dots in the middle panel of Fig. 6) have an  
 531 increasing, with the simulated streamflow magnitude, negative mean. This effect can also be observed in PP-Plots of Fig. 4,



532 where for the GL++ inference the curvature is pronounced and far away from an “S-shape”. Therefore, the GL++ error model  
 533 infers the bivariate distribution  $p(e, q_s)$  considering the position of all its error conditional distributions centered at zero  
 534 (i.e. with a null conditional bias) and given this, it estimates its shape properties (e.g. its variance). The result is a poor  
 535 performance of the variance model given by Eq. (6), which is not able to fit the correct amplitude of the bivariate distribution  
 536 without considering correctly its location. In short, in presence of a non-constant bias, a variance model on its own is not able  
 537 to correctly model the bivariate distribution of the simulated discharges and its errors. The expression of the TVL in Eq. (7)  
 538 summarizes this idea, since it considers both conditional bias and conditional variance terms to obtain the total (marginal)  
 539 variance of the hydrological model errors.



540 **Figure 6. Test of GL++ error model hypotheses for CRR (left column) and GR4J (right column) models. In top panels, the evaluation**  
 541 **of the fitting to the SEP distribution, of the observed standard innovations. In middle panels, the homoscedasticity evaluation for the**  
 542 **observed standard errors (red dots and red line) and for the observed standard innovations (black dots and grey line). In bottom**  
 543 **panels, the independence assessment for observed errors and for the innovations, through the autocorrelation function (the 95%**  
 544 **Anderson’s limits in solid black lines).**

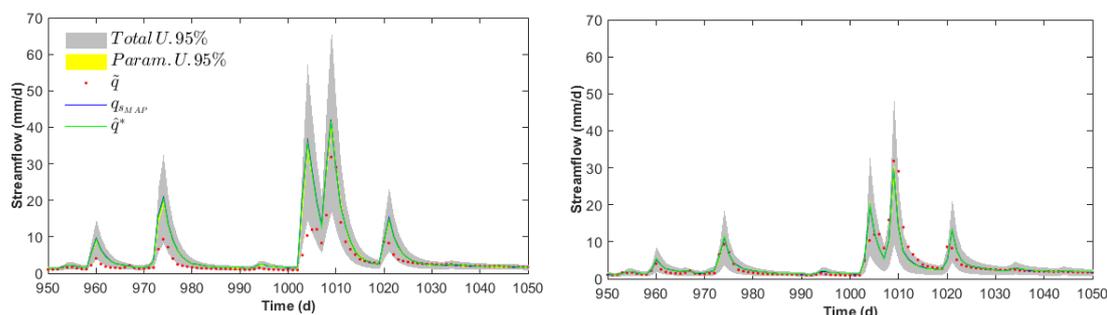
545 In view of a better fulfillment of the error model hypotheses (compared with SLS and WLS), it is expected that the GL++  
 546 parameter estimation could be less biased than the corresponding to those classical schemes of inference (Schoups and Vrugt,  
 547 2010). In fact, this is the reason for the poor performance shown by the biased prediction of the hydrological model: the most  
 548 plausible parameter set, for both hydrological and error models, brings out (in form of a prediction bias) the deficiencies in  
 549 the hydrological model and/or in the input data.

550 Concerning the above mentioned overpredictive bias, Table 1 shows some of its effects. There is deterioration in the NSE  
 551 index and VE for the mean prediction, with respect SLS and WLS inferences, for both hydrological models. It must be also



552 underlined the great difference between the performance of the two models which arises with GL++ inference. In contrast,  
553 with SLS and WLS inferences, the performance of both hydrological models resulted similar. The CRR performance has  
554 suffered an important deterioration, whereas the GR4J performance is still acceptable. The meaning of all these results is that  
555 GL++ inference forces a more realistic performance of the hydrological models than in SLS or WLS inferences: and this is  
556 indicating that GR4J is able to model this case study with a better performance than the CRR model. In other words, for this  
557 case study GR4J with only 4 parameters is less affected (it has a less biased mean prediction) by the structural and/or data  
558 deficiencies than the CRR model, which have 7 parameters.

559 The reliability assessment of the PD, through the PP-Plots in Fig. 4, shows for CRR a large overpredictive bias, much bigger  
560 than the GR4J one, confirming the previously explained mean prediction performance for both models. Furthermore, the  
561 overprediction of both hydrological models (bigger for CRR) can also be observed in Fig. 7, where the mean prediction  
562 (green solid lines) is more often above observations (red dots). In relation to the uncertainty assessment, PP-Plots in Fig. 4  
563 show its correctness for both models. Besides, Fig. 7 shows the differences between CRR and GR4J in the appearance of the  
564 95% total uncertainty band: these differences are more visible in the streamflow peaks, where the overestimation in CRR also  
565 induces the greatest widening of this uncertainty band.



566 **Figure 7. Predictive performance and 95% uncertainty bands (parameter and total) with GL++ inference on CRR (left) and GR4J**  
567 **(right) models (parameter uncertainty band is hardly appreciable).**

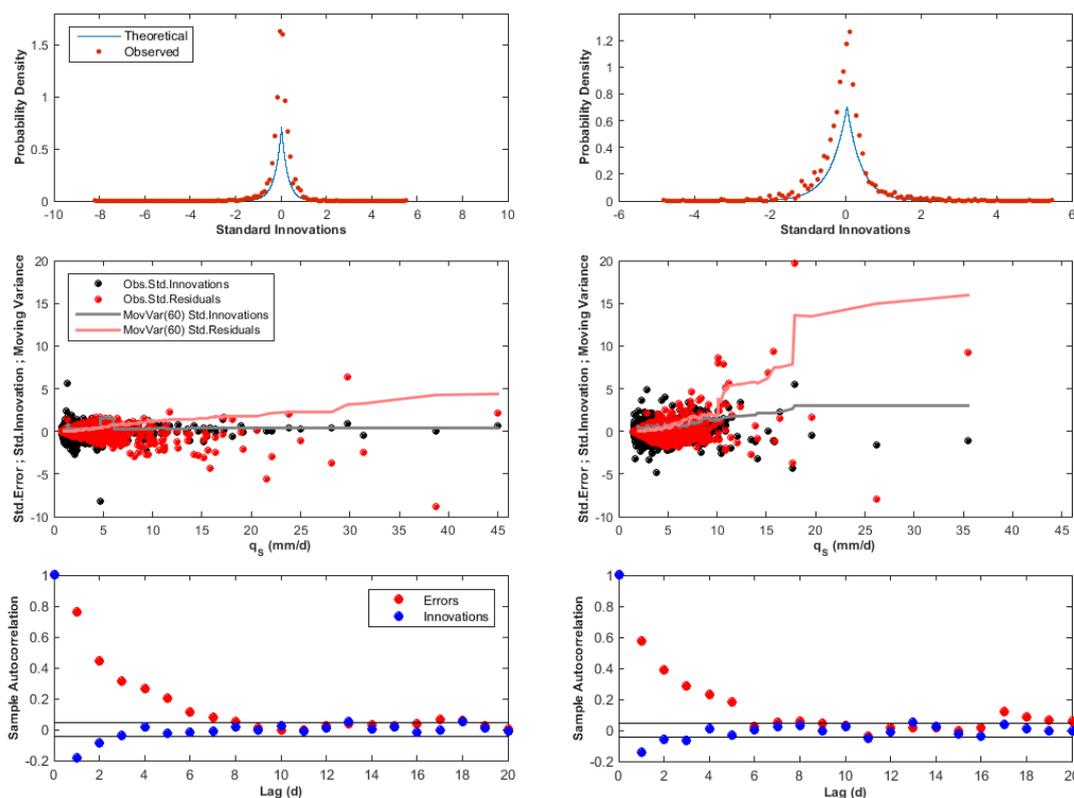
#### 568 5.4 GL++Bias inference

569 As it has been underlined, GL++ inferences introduce a bias in the PD of both hydrological models. The bias model in  
570 GL++Bias inference, which was described in section 2 by Eq. (5), considers that error bias depends on the simulated  
571 streamflow, as in the variance model, but with the flexibility of a different behavior for low and high flows. Figure 10  
572 represents, for both CRR and GR4J hydrological models, a scatter of the errors jointly with their inferred variance and bias  
573 models. In this figure, it can be noticed the different behavior of the bias depending on the hydrological model. On the one  
574 hand, inference on CRR distinguishes (see zoomed window in Fig. 10) a different bias slope between low flows and high  
575 flows, with a threshold  $q_o=1.96$  mm/d. On the other hand, GR4J practically does not make this distinction and the inferred  
576 threshold  $q_o=1.57$  mm/d is very close to the minimum simulated streamflow. In any case, Fig. 10 shows the good fitting of  
577 the used deterministic simple bias model: the inferred lines (blue dashed), representing the error conditional means, are  
578 placed at the central position of the errors scatter.

579 The inference with the GL++Bias error model yields a reasonably good fulfillment of the hypotheses about the errors. In any  
580 case, better fulfillment than previously analyzed inferences, for the case study. There are remaining problems which are  
581 described in what follows. This inference produces for both hydrological model (see Fig. 8, top panel) a pdf for the  
582 innovations with a high density for the zero and near-zero values; namely, a more kurtotic observed innovations distribution  
583 than the inferred SEP. Besides, and commonly to the previously performed GL++ inferences, the innovations autocorrelation  
584 for the first lags is not completely removed, although its values are relatively small (see Fig. 8, bottom panel). In respect to  
585 the variance stabilization, which had a poor performance in GL++, it can be observed in middle panel of Fig. 8 that the  
586 consideration of the bias model allows a good stabilization of the innovations variance, with a near-horizontal moving



587 variance line in the CRR case, whereas for the GR4J the stabilization is also acceptable, albeit the line is not completely  
 588 horizontal.



589 **Figure 8. Test of GL++Bias error model hypotheses for CRR (left column) and GR4J (right column) models. In top panels, the**  
 590 **evaluation of the fitting to the SEP distribution, of the observed standard innovations. In middle panels, the homoscedasticity**  
 591 **evaluation for the observed standard errors (red dots and red line) and for the observed standard innovations (black dots and grey**  
 592 **line). In bottom panels, the independence assessment for observed errors and for the innovations, through the autocorrelation**  
 593 **function (the 95% Anderson's limits in solid black lines).**

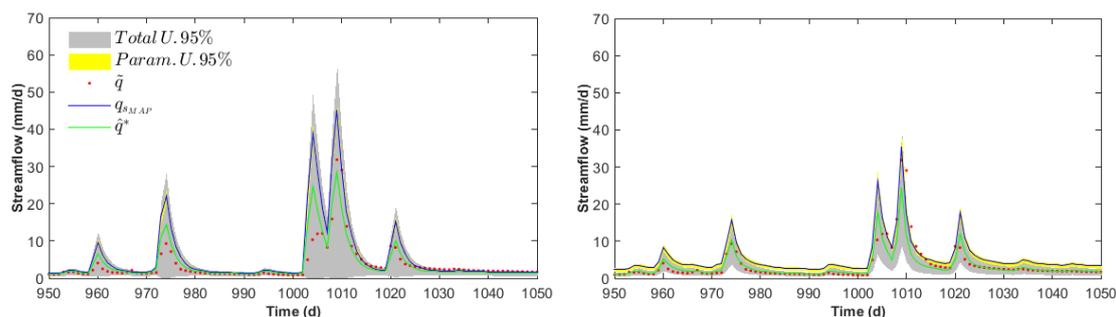
594 The performance of the GL++Bias error model combined with the CRR hydrological model is very different from the  
 595 performance with the GR4J hydrological model. As it was explained in the GL++ inference, the CRR model showed a more  
 596 biased prediction than GR4J. In GL++Bias, the proposed bias model is able to considerably improve the performance of the  
 597 mean prediction in CRR (green line in Fig. 9), where prediction bias is practically removed. In the case of GR4J, the bias  
 598 model produces an improvement of the prediction, which is more modest: a small overpredictive bias still remains with  
 599 GR4J, after the bias model inclusion.

600 Table 1 shows that CRR model with GL++Bias obtains for the mean prediction a NSE=0.76 and a VE=0.0%, whereas GR4J  
 601 yields a NSE=0.80 and also a VE=0.0%. The null error in volume, confirms a fair performance (considered in average over  
 602 all the calibration period) of the bias model. In general terms, the comparison between GL++ and GL++Bias shows that the  
 603 CRR mean prediction performance has improved notably (from NSE=0.25 to NSE=0.76), but for GR4J the performance is  
 604 nearly the same in both inferences (slightly decreasing from NSE=0.82 to NSE=0.80). We can conclude that the prediction  
 605 bias which arises from the GR4J model with GL++ is small, but with a structure too complex to be well-reproduced by the  
 606 bias model hypothesized by GL++Bias. On the contrary, most of the large bias which arises from the CRR model with GL++  
 607 can be corrected with the bias model given by GL++Bias.



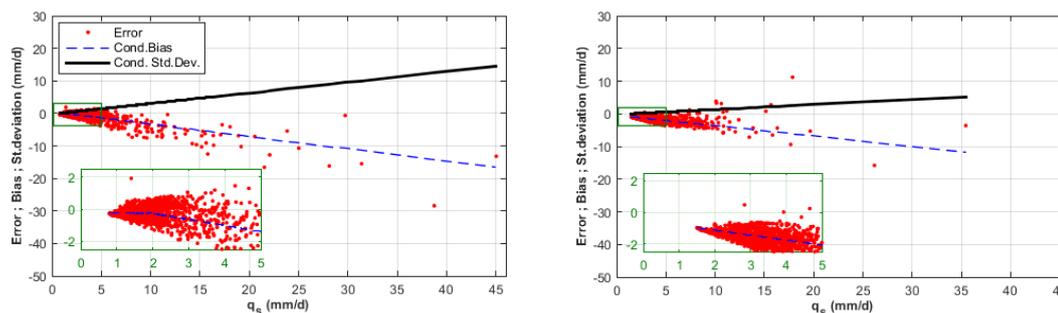
608 Regarding the predictive uncertainty assessment, both GL++Bias inferences yield a reduction of the slope parameter in the  
 609 variance model, in comparison with the corresponding GL++ inference (Tables 2 and 3, and black lines in Figs. 10 and 11).  
 610 This reduction is larger for GR4J (from 0.35 to 0.15) than for CRR hydrological model (from 0.42 to 0.33). Due to this  
 611 reason, the PD width for GR4J model is smaller than for CRR, which can be observed comparing the two 95% uncertainty  
 612 bands in Fig. 9. That is to say, when GL++Bias error model is used, predictions with CRR are more uncertain than with  
 613 GR4J, but are also less biased. The PP-Plots (Fig. 4) exhibit the good performance of the PD for both hydrological models:  
 614 even the CRR model with the GL++Bias error model shows a near-perfect fit to the 1:1 line.

615 Furthermore, looking at right panel of the Fig. 9, it is important to realize that the GL++Bias inference for GR4J model is the  
 616 only inference that exhibits a significant contribution of parameter uncertainty to the total predictive uncertainty. This  
 617 contribution seems to be underestimated in all the other performed inferences.



618 **Figure 9. Predictive performance and 95% uncertainty bands (parameter and total) with GL++Bias inference on CRR (left) and**  
 619 **GR4J (right) hydrological models.**

620



621 **Figure 10. GL++Bias error model features for CRR (left) and GR4J (right) hydrological models: inferred conditional variance**  
 622 **(black line) and inferred conditional bias (dashed blue line) with the error versus  $q_s$  scatter plot. The zoom window shows the**  
 623 **performance of the inferred bias model for low flows.**

624

625

626

627

628



## 629 6. Discussion

### 630 6.1 Comparison between enforcing (TL) or not enforcing (NTL) the Total Variance Law

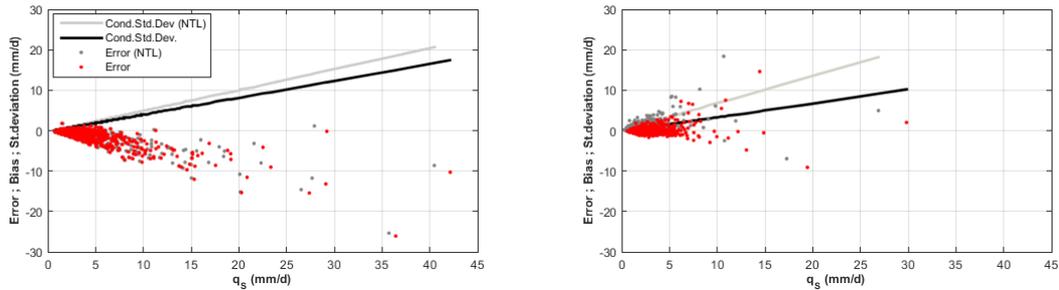
631 In this research, we are assuming hydrological model errors are part of a joint (bivariate) probability distribution between  
 632 errors and simulated discharges at the catchment outlet. The conditional and marginal distributions belonging to a joint  
 633 distribution must fulfill the TLs. As it was exposed in previous sections: i) SLS hypotheses “automatically” induce the TLs  
 634 fulfillment; ii) when error statistical features are hypothesized as non-constant, and they are expressed with parametric  
 635 functional forms as in Eqs. (5) and (6), the inference problem must be constrained to ensure the TLs fulfillment; iii) if the  
 636 error model assumes no error bias, TEL is supposed to be “automatically” fulfilled. When TLs are not fulfilled, the inferred  
 637 errors and the corresponding state variable (streamflow in this paper) cannot belong to the same bivariate joint distribution  
 638 and therefore, the predictive uncertainty problem to solve is statistically incorrect.

639 This incorrectness generates problems, mainly related with spurious parameter interactions, affecting the inference results  
 640 and making them unsuitable and possibly non-robust (Evin et al., 2014). This section will demonstrate that not enforcing the  
 641 TLs is, at least, one of the most important causes of these problems. This demonstration has been done for the WLS and  
 642 GL++ error models comparing the results with and without (NTL) the TLs enforcement. Since these two error models assume  
 643 no error bias, only TVL must be enforced in the joint inference.

644 In the WLS comparison, we have found that results with and without TVL enforcement are very similar. This occurs with the  
 645 two used hydrological models. The inferred hydrological parameters result similar (see Tables 2 and 3), so the hydrological  
 646 model behavior is practically the same and the likelihoods are also similar (not shown). Concerning the error variance model  
 647 parameters, they are not exactly equal, as Tables 2 and 3 show, but the difference seems to be negligible (e.g. for CRR model  
 648  $K = 0.24$  in NTL case and  $K = 0.27$  in TL case). Therefore, it seems that WLS inference would also be able to fulfill the  
 649 TLs (as SLS), without the need of enforce them. However, this is not the case. WLS generally will not fulfill the TVL if this  
 650 is not enforced. But this is not an important issue in the case of the WLS inference, as it is explained in the following. Let  
 651  $\{\alpha, \kappa\}_{ML, TL}$  be the set of variance model inferred parameters that fulfills both the maximum likelihood and the TLs criteria.

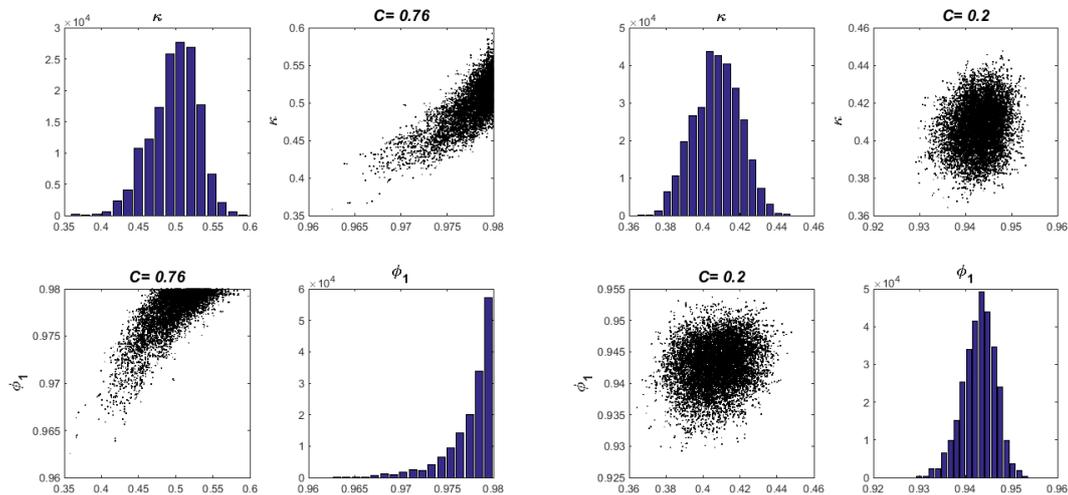
652 It can be demonstrated that with WLS, any other set of parameters such that  $\{\alpha, \kappa\}_{ML} = \mathbf{C} \{\alpha, \kappa\}_{ML, TL}$ , where  $\mathbf{C} > \mathbf{0}$  is a  
 653 proportionality factor, will have the same maximum likelihood value, but it will not fulfill the TVL. Hence, with the WLS  
 654 error model, the same hydrological parameter estimation is inferred, as well as similar uncertainty bands are obtained, using  
 655  $\{\alpha, \kappa\}_{ML, TL}$  or using any other  $\mathbf{C} \{\alpha, \kappa\}_{ML, TL}$ . This means that WLS, in NTL case, is a bad-posed problem to estimate a  
 656 unique identifiable set of variance model parameters. In the case study, for the CRR hydrological model, a proportionality  
 657 factor of  $\mathbf{C} = 1.125$  has been obtained, as it can be checked in Table 2.

658 Regarding the comparison of the GL++ joint inferences, with and without the TLs enforcement, this research demonstrates  
 659 that when the error model considers both the autocorrelation (AR) and the heteroscedasticity models, not enforcing the TLs  
 660 has a significant effect on the result of the inference (unlike to what has been resulted for WLS error model). Figure 11 shows  
 661 the error scatter plots as well as the inferred linear variance models for the GL++ inferences with and without the TVL  
 662 enforcement. In these cases, the effects of not considering the TLs with GR4J (right panel) are always substantially larger  
 663 than with CRR (left panel). These effects are the following: i) an increment of the slope in the linear variance model (see Fig.  
 664 11), from TL enforcement (black lines) to NTL (gray lines); ii) a change in the hydrological parameters, significant with  
 665 GR4J (Table 3) and small with CRR hydrological model (Table 2); iii) an important erroneous uncertainty overestimation, as  
 666 can be observed by the comparison between Figs. 7 (TL case) and 13 (NTL case), for the CRR model (left panels) and for  
 667 GR4J (right panels); iv) changes in the PD or even radically different PDs, as it is shown for the GR4J model in the  
 668 corresponding PP-Plot of the right panel in Fig. 4, where with GL++ there is a moderate overprediction, and with GL++NTL  
 669 a strong underprediction; v) in NTL inferences it is observed a high correlation (as a spurious parameter interaction, which it  
 670 does not exist when TLs are applied) between the slope parameter of the error standard deviation and the autocorrelation  
 671 parameter (Fig. 12). This last feature was also reported in Evin et al. (2013, 2014).

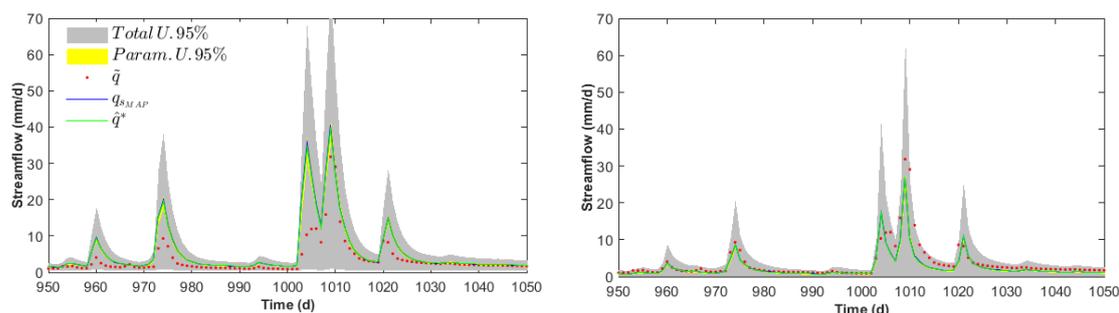


672 **Figure 11. GL++ and GL++NTL error model features for CRR (left) and GR4J (right) hydrological models. TVL enforcement case:**  
 673 **inferred variance model (black line) and error versus  $q_s$  scatter (red). NTL case: inferred variance model (grey line) and error versus**  
 674  **$q_s$  scatter (grey).**

675 We have numerically tested in the case study the fulfillment or not of the TLs. Because WLS and GL++ assume no error bias,  
 676 TVL in Eq. (7) simplifies to Eq. (21). For the CRR-GL++ inference the calculated error marginal variance value (left term in  
 677 Eq. (21)) is  $2.25 \text{ mm}^2/\text{d}^2$ , which exactly matches the value for the mean of the conditional variances (right term in Eq. (21)).  
 678 However, in the CRR-GL++NTL case, the calculated error marginal variance is  $1.97 \text{ mm}^2/\text{d}^2$ , which is very different from  
 679 the mean of the conditional variances ( $3.31 \text{ mm}^2/\text{d}^2$ ). Therefore, in the NTL case, as it was expected, there is an imbalance  
 680 between the marginal and conditional error variances.



681 **Figure 12. Posterior distribution of the slope parameter of the variance model, posterior distribution of the autocorrelation**  
 682 **parameter of the AR(1) model and scatter plot between them, for inferences with CRR hydrological model: left panels with**  
 683 **GL++NTL error model; right panels with GL++ error model.**



684 **Figure 13. Predictive performance and 95% uncertainty bands (parameter and total) with GL++ NTL inference on CRR (left) and**  
 685 **GR4J (right) models (parameter uncertainty band is hardly appreciable).**

## 686 **6.2 Hydrological parameter estimates depending on the error model**

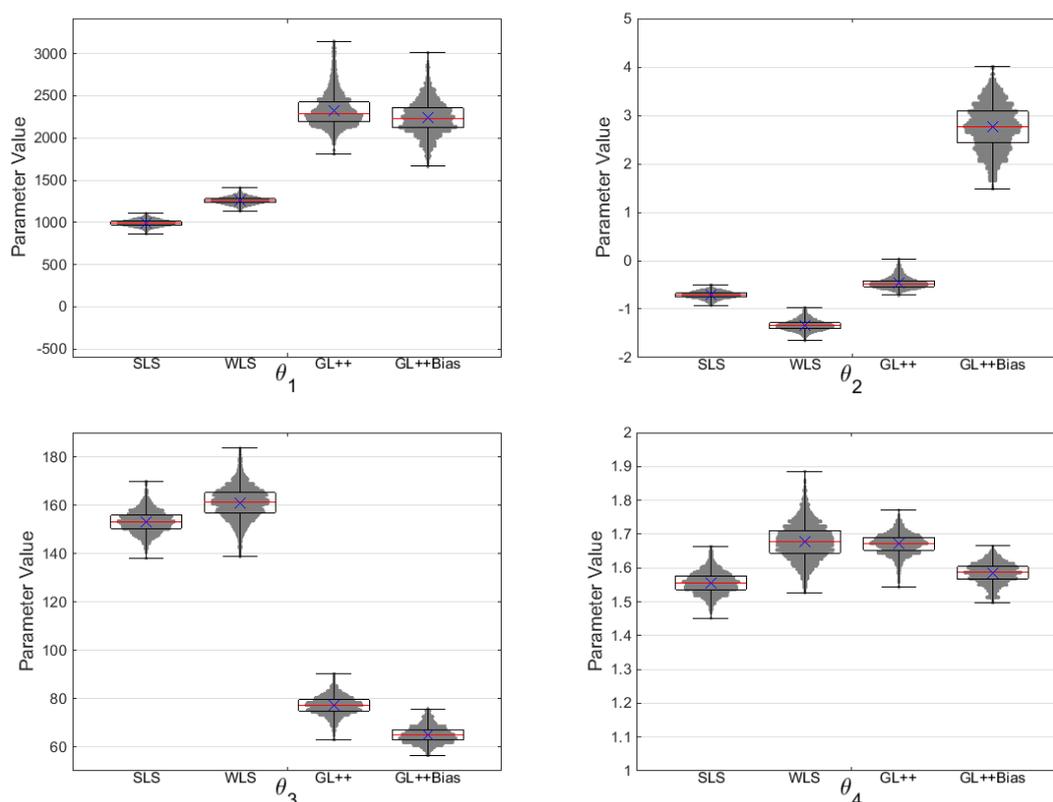
687 It has already stated in previous sections that an incorrect error model (actually an incorrect likelihood function) can yield  
 688 apparently very good mean predictions in calibration, but with biased hydrological parameters and an incorrect uncertainty  
 689 assessment. The paradigm of this idea is the SLS error model, widely used in hydrological modeling. On the contrary, a  
 690 correct consideration of the errors in the inference yields more suitable parameter estimation and correct hydrological  
 691 predictions, showing the model deficiencies in the form of a biased model results and an acceptable uncertainty band, even  
 692 with a poor fitting to the observed data. The analyzed case study has at least a known structural problem in both models, the  
 693 neglected snow processes, and for sure other problems not yet identified in the model or in the data. All these circumstances  
 694 predetermine the errors structure, departing it from the SLS hypotheses of zero bias, normality, independence and  
 695 homoscedasticity. Previous sections have shown how the predictions and their uncertainty evolve as a function of the  
 696 hypothesized error model, from the SLS to the GL++Bias. This section will show the “other side of the coin”: how the  
 697 parameters change with the error model.

698 To illustrate the effect of the error model on the inferred parameters, it has been chosen the GR4J model, since it is  
 699 parsimonious and all its four parameters have been perfectly identifiable in all performed inferences, which did not occur  
 700 with the CRR model. Figure 14 presents the four parameter marginal posteriors of GR4J hydrological model when they are  
 701 jointly inferred with the SLS, WLS, GL++ and GL++Bias error models. There are important differences between the  
 702 marginal posterior distributions, especially for parameters  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$ . Another important characteristic to analyze is the  
 703 parameter uncertainty: for parameters  $\theta_1$  and  $\theta_2$  it tends to increase with the more sophisticated error models, whereas for  
 704 parameters  $\theta_3$  and  $\theta_4$  it remains similar. Concerning the corresponding CVs (Table 3), systematically SLS gives smaller  
 705 values.

706 But the most paradigmatic and didactic proof of the necessity of making the parameter inference with a proper error model is  
 707 the case of the GR4J parameter  $\theta_2$  (the groundwater exchange coefficient). Figure 14 shows how the estimated  $\theta_2$  suffers  
 708 important changes in its value (and even in the sign!) depending on the inference. This has a deep consequence in the model  
 709 conceptualization of the analyzed basin. For GR4J, a negative value of  $\theta_2$  means that a deep (or regional) aquifer receives  
 710 water from the basin. In contrast, a positive value of  $\theta_2$  means that this regional aquifer feeds the basin with groundwater  
 711 (whose origin is out of the basin). On the basis of our knowledge about the geology of FB basin, it does not have a deep  
 712 aquifer with which exchange water. Hence a reasonable value for  $\theta_2$  should be zero. I.e. the water balance of the basin must  
 713 be closed without losses or contributions different to actual evapotranspiration or precipitation respectively. In absence of a  
 714 regional aquifer, different values to zero are due to model and/or input deficiencies and they do not have any hydrological  
 715 meaning, although they enable to close the water balance. In our case study, SLS inferred MAP value is  $\theta_2 = -0.71$  mm, WLS  
 716 MAP value is still more negative, whereas with the GL++, we get a MAP value of  $\theta_2 = -0.55$  mm. All three inferences without  
 717 bias need dewater the basin (negative values for  $\theta_2$ ). Nevertheless, the most plausible inferred value for  $\theta_2$  (the closest to  
 718 zero) corresponds to GL++, the most correct among these three error models.



719 Regarding GL++Bias inference, its MAP value is  $\theta_2=2.89$  mm. According to the previously exposed, in our case study this is  
 720 an implausible value for this parameter (it is the most different from zero). Another important feature to note is that the  
 721 uncertainty of  $\theta_2$  with GL++Bias is the highest of all hydrological parameters and performed inferences, as it is shown in Fig.  
 722 14 and reflected in the CV values for this parameter in Table 3. This high uncertainty is the main responsible of the thick  
 723 predictive uncertainty band due to the parameters, which was shown in Fig. 9 (right panel). In short, neglecting the bias  
 724 modeling or considering an incorrect error bias model produces a biased estimation of parameter  $\theta_2$ . In general, incorrect  
 725 error models will produce biased parameter estimates.



726 **Figure 14. Evolution of the posterior distribution (box-plots) of the four GR4J model parameters with the four hypothesized error**  
 727 **models.**

### 728 6.3 Performance of the different inferences

729 There are several criteria in which we could rely to state why or why not an inference should be considered as a suitable  
 730 inference, given a hydrological model. The first general criterion should be the degree of fulfilment of each error model  
 731 hypotheses. In other words, assessing to which extent the inferred errors are unbiased, homoscedastic and independent, and  
 732 also show a good fitting to the hypothesized distribution. This analysis has been exhaustively done in section 5.

733 A second general criterion (which could be included into the previous one) is how accurate are the hypothesized error bias  
 734 and variance models. This accuracy can be easily checked by testing if the inferred standardized errors  $\eta$ , defined in Eq. (8),  
 735 have zero mean and unit variance. The mean and standard deviation of the standardized errors ( $\bar{\eta}$  and  $S_{\eta}$  respectively) are



736 shown in Table 1 for the case study. This criterion is focusing on the goodness of the error standardization (hereinafter  
737 GoES). The meaning of getting other different values to the theoretical ones (zero or one respectively) is that we are shifting  
738 and scaling the errors but not making their correct standardization; namely, the hypotheses for the error bias and/or variance  
739 models are unsuitable. It is important to realize that the fulfillment of both TLs is a necessary but not sufficient condition to  
740 obtain standardized errors with statistics satisfying or near-satisfying their theoretical values: i.e. the suitability of the error  
741 model components is also a necessary condition. Previous works (Evin et al., 2013, 2014; Scharnagl et al., 2015), where TLs  
742 were not enforced, considered that the standardized errors variance  $S_{\eta}^2$  is, in any case, equal to one. However, inferred  $S_{\eta}^2$   
743 can be very different from the unit when TLs are far to be fulfilled. As an example with the GR4J hydrological model, this  
744 research got a value of  $S_{\eta}=0.2$  for the GL++NTL inference, instead of the more correct value of  $S_{\eta}=0.95$  for the  
745 corresponding GL++ inference (see Table 1).

746 The third general criterion to distinguish a fair inference is based on the general reliability of the PD. This can be visually  
747 assessed by means of the PP-Plots (as was done in section 5). Moreover the reliability of the PD can be quantified through  
748 the reliability and the resolution (or degree of certainty) indexes, defined as in Renard et al. (2010) and shown in Table 1.  
749 Generally, an inference with a proper GoES also has high reliability indexes (as occurs with the CRR-GL++Bias inference),  
750 but the opposite is not necessarily true (as occurs with the GR4J-GL++Bias inference). Given several inferences with a  
751 similar PD reliability index, those with the lowest prediction uncertainty, namely with the highest resolution, are preferable.

752 Finally, as the fourth general criterion, we should check and discard those inferences which exhibit parameter values whose  
753 meaning may not correspond to the knowledge we have about the basin behavior. This was already analyzed for one  
754 hydrological model in section 6.2.

755 Therefore, only when all previous criteria are acceptably satisfied, we could pay attention to the performance of the expected  
756 prediction (tested by its NSE and VE indexes shown in Table 1). That is to say, prediction performance of the hydrological  
757 model is a consequence (rather than a selection criterion) and it should be subordinated to the fulfillment of all previously  
758 exposed criteria. It is important to realize that, in this way, it will be obtained a set of parameter estimates which are expected  
759 to be more robust than those obtained with the classical least-squared based optimization methods.

760 These criteria have been applied to all the implemented inferences. Regarding SLS, the conclusion is direct: fails to meet the  
761 most of the requirements, but it has the useless property of having generally good performances of the expected prediction in  
762 calibration. Concerning WLS, this error model only fulfills with the error homoscedasticity hypothesis, and generally also  
763 exhibits good reliabilities of the PD in calibration. However, since WLS does not remove the errors serial dependence  
764 structure, it cannot be considered a correct error model when this problem appears. With respect to the GL++ and GL++Bias  
765 error models, conclusions are different depending on the hydrological model applied in the joint inference. But, a common  
766 behavior for these inferences is an acceptable fulfillment of the error model hypotheses, which is an important advance over  
767 the classical SLS and WLS error models.

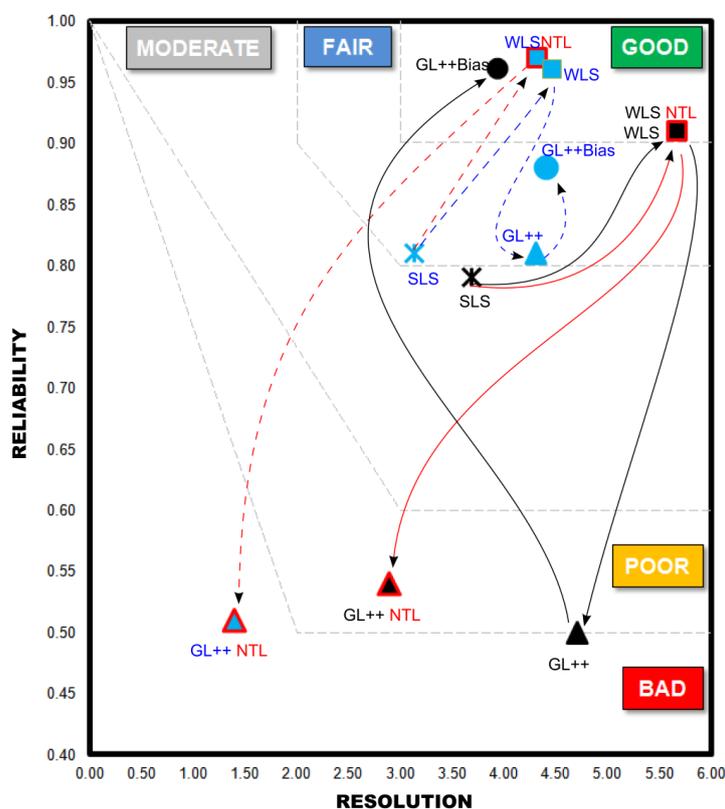
768 To complete the discussion about both GL++ and GL++Bias error models, we will focus on Fig. 15. This figure shows a  
769 reliability-resolution diagram where are represented all inferences. The reliability-resolution diagram is useful for comparing  
770 several inferences. It is important to note that the limits between classification zones in the diagram are subjective, based on  
771 our experience, and they are provided only for guidance. The lines in the diagram are the trajectories from the simplest to the  
772 more advanced error model, starting with SLS as the reference inference. Furthermore, the NTL inferences where TLs have  
773 not been enforced are also represented with red trajectories. The main aspects to be considered in relation with Fig. 15 are the  
774 following:

- 775 i) As it has already been exposed, this diagram shows how WLS improves both reliability and resolution in  
776 relation to SLS, yielding a very good PD performance in calibration. Figure 15 also shows how WLS inferences  
777 yield similar points in the diagram between the NTL and TL cases, as theoretically expected (see section 6.1).
- 778 ii) GL++ inferences deteriorate the PD reliability in comparison with WLS. The evolution from WLS to GL++,  
779 which involves the application of an autoregressive model on the errors, is able to “convert” any of the



780 parameter bias yielded by WLS into a bias in the prediction, but with the worsening of the PD reliability. It is  
 781 expected that, the more the PD reliability difference between WLS and GL++ is, the more the parameter bias in  
 782 WLS and the prediction bias in the GL++ are. According to this, it can be noticed (Fig. 15) that distance  
 783 between WLS and GL++ in GR4J is much smaller than in CRR, indicating that GR4J has less structural  
 784 deficiencies than CRR. I.e. GR4J is a better model for this case study.

785



786 **Figure 15. Reliability-Resolution diagram. Comparison of the PD performance for all inferences with the TLs enforcement: with**  
 787 **CRR (continuous black trajectories) and with GR4J (dashed blue trajectories). Consequences on the PD when TLs are not enforced**  
 788 **(NTL cases): with CRR model (continuous red trajectories) and with GR4J model (dashed red trajectories).**

- 789 iii) The enforcement of TLs on GL++ inferences improves the performance of the PD in resolution and/or  
 790 reliability in respect to the corresponding GL++NTL inferences. In our case study and for the CRR model, the  
 791 improvement is mainly on the resolution, whereas with the GR4J model the inference with the enforcement of  
 792 TLs yields a great improvement on both resolution and reliability indicators. In fact, for both hydrologic  
 793 models, we can see that the GL++NTL inferences show a PD performance even worse (in both resolution and  
 794 reliability) than the corresponding SLS inferences.
- 795 iv) The addition of a bias model to GL++ corrects in some degree the worsening in the reliability produced in the  
 796 transition from WLS to GL++ for both hydrological models. The achieved reliability in the GL++Bias inference  
 797 depends on the suitability of all the hypothesized error model components (bias, variance and dependence  
 798 models) to reproduce the complexity of the error structure. The second criterion could be a good benchmark  
 799 to test this suitability. So, for the CRR model, the GL++Bias inference yields the best reliability (see the good



800 fitting to the diagonal of the PP-Plot in Fig. 4) and it also produces the best GoES statistics, as shown in Table  
801 1. Therefore, the GL++bias inference seems to be the most suitable for the CRR model. Focusing on the GR4J  
802 model, GL++Bias inference slightly improves the reliability and the resolution, in relation to the corresponding  
803 GL++ inference. However, GL++Bias obtains GoES statistics ( $S_\eta=1.5$  and  $\bar{\eta}=-0.28$ ) which are moderately  
804 worse than for the GL++ inference ( $S_\eta=0.95$  and  $\bar{\eta}=-0.26$ ). We can conclude that the inferred simple error  
805 bias model is not as suitable for GR4J as it seems to be for CRR.

## 806 7. Conclusions

807 This paper has addressed the challenging problem of jointly estimate hydrological and error model parameters in a Bayesian  
808 framework, trying to solve some of the problems found in previous related researches, as in the second case study of Schoups  
809 and Vrugt (2010) as well as in Evin et al. (2014), among others.

810 Firstly, we have developed a new general formal likelihood function for parameters and predictive inference of hydrological  
811 models with correlated, biased, heteroscedastic, and/or non-Gaussian errors, which is based on the general likelihood  
812 function developed in Schoups and Vrugt (2010), but with the methodological modifications proposed by Evin et al. (2013).

813 Secondly, the detection of pitfalls in the joint inference methodology was not completed with the Evin et al. (2013) paper. In  
814 that paper and all previous related ones, the non-constant error variance and the error serial dependence were jointly modeled  
815 without considering that the errors and the simulated streamflow conform a joint distribution  $p(e, q_s)$ . Therefore, they also

816 did not consider any relationship between the error conditional (on  $q_s$ ) distributions and their associated error marginal  
817 distribution. The main finding of this paper is that a joint inference, to be correct, must take into account the joint distribution  
818  $p(e, q_s)$  and consequently the relationship between its marginal and conditional distributions. This relation is defined by  
819 two general statistical expressions called the Total Laws (TLs): the Total Expectation and the Total Variance Laws. An  
820 important consequence of the TLs enforcement is the reduction of the degrees of freedom in the inference problem: namely,  
821 the reduction of the parameter space dimension. For the case of considering more than one state variable of interest, any joint  
822 inference must take into account the joint probability distribution of all variables to be predicted and its corresponding errors.  
823 This hyper-joint distribution must also fulfill TLs.

824 In the presented case study, the Bayesian joint inference has been performed through the application of several inference  
825 models, as the known SLS or WLS and the new GL++ and GL++Bias. The Differential Evolution Adaptive Metropolis  
826 algorithm Dream-ZS (Schoups and Vrugt, 2010; Laloy and Vrugt, 2012) has been used to sample the posterior distribution of  
827 the hydrological and error model parameters. The inferences were carried out on two lumped hydrological models (the CRR  
828 model and the GR4J model) which were forced with daily hydrometeorological data from a basin of the MOPEX project: the  
829 French Broad basin, a wet basin of 2448 km<sup>2</sup>. The main conclusions derived from the case study analysis are summarized in  
830 the following paragraphs.

831 Surplus non-random information is accumulated in the errors when a hydrological model is not able to process the inputs in a  
832 correct way to reproduce the state variable observations. This model incapacity is due to deficiencies (misspecifications) in  
833 the model structure and/or in the observed input/output data reliability. The non-random errors show several symptoms (e.g.  
834 marginal and/or conditional non-zero-mean distributions, autocorrelation) which alert us about the presence of problems in  
835 the model and/or data. In presence of these problems and, if we apply the “true” but generally unknown parameter set,  
836 hydrological model simulations could be far from the observed values. Having this in mind, it is important to note that only  
837 when the jointly inferred errors fulfill the hypotheses of the error model, the hydrological model shows the most realistic  
838 prediction and the less biased parameter estimates. If we do not want or cannot improve the hydrological model and data  
839 reliability, the error model must take into account (in an aggregated way) all these deficiencies. That is the reason why GL++  
840 inference has brought to light the deficiencies of both hydrological models, as a bias in the prediction. Moreover, GL++ has  
841 enabled to determine that, for the case study, GR4J is a better model (because shows a less biased prediction) than CRR,



842 whereas the results of the SLS inference could suggest the opposite. In short, structural or data problems necessarily  
 843 undermine either the reliability of the inferred parameters or the performance of the model results, depending on the chosen  
 844 error model.

845 The non-fulfillment of the TLs is statistically incorrect when we are modeling a joint pdf through the definition of its  
 846 conditional distributions. Only simple error models, as SLS, do not explicitly need the TLs implementation. Non-fulfillment  
 847 of TLs produces incorrect parameter estimates and little reliable predictive distributions, as it was shown in NTL cases.  
 848 Probably this is also the main cause of the spurious parameter interactions found in *Evin et al.* (2014) and shown in Fig. 12  
 849 for our case study. However, the TLs enforcement is a necessary but not sufficient condition for the correct parameter  
 850 estimation and predictive uncertainty assessment: a misspecification or incorrect hypothesis about the errors statistical  
 851 features can also yield biased estimates and little reliable predictive distributions, as it was also reported by Reichert and  
 852 Schuwirth (2012). For example, we have demonstrated that ignoring the modeling of the error biases, when they exist,  
 853 produces biased predictions. This has been shown with the GL++ inferences yielding biased predictions with both  
 854 hydrological models. Besides, ignoring the bias or modeling it incorrectly produces biased parameters regarding the “true”  
 855 parameter set. This was clearly shown for the parameter  $\theta_2$  of the GR4J model.

856 A Bayesian formal approach to address the inference of hydrological models, as proposed in this paper, breaks away with the  
 857 paradigm of parameter estimation methods based on minimizing summary statistics of model errors. In this context, this  
 858 research has followed several selection criteria which are useful to discard or accept different inferences. As summary and in  
 859 sequential order: i) the fulfillment of the error model hypotheses, including the adequacy of the bias and variance models to  
 860 the inferred errors; ii) the general reliability of the PD; and iii) the plausibility of the parameter values in the context of the  
 861 hydrological knowledge about the modeled basin. However, given the hydrological model, the resulting performance of the  
 862 prediction, the reliability of its predictive uncertainty, as well as the robustness of the parameter estimates, will be exclusively  
 863 conditioned by the first criterion: the degree in which errors fulfill the error model hypotheses.

## 864 8. Acknowledgments

865 This research has received funding from the Spanish Ministry of Economy and Competitiveness through the research projects  
 866 SCARCE CONSOLIDER (CSD2009-00065), ECO-TETIS (CGL2011-28776-C02-01) and TETISMED (CGL2014-58127-  
 867 C3-3-R). We also wish to acknowledge Dr. Ezio Todini and Dr. Jasper Vrugt for constructive feedback that has helped to  
 868 improve the manuscript. Dr. Vrugt is also acknowledged, for providing the code for the DREAM-ZS algorithm. The data  
 869 used in this research can be downloaded from [http://hydrology.nws.noaa.gov/pub/gcip/mopex/US\\_Data/](http://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/).

## 870 Appendix A

871 This appendix shows how to obtain the general likelihood function given by Eq. (17). This new likelihood function evolves  
 872 from the original version in Schoups and Vrugt (2010), implementing the methodological correction proposed in Evin et al.  
 873 (2013) according to which errors must be studentized before applying an autoregressive error model on them.

874 From the viewpoint of Bayesian statistics, not only variable observations but also model parameters are considered as random  
 875 variables. Therefore, a density function of multivariate joint probability (hyper-surface) with the observations and the  
 876 parameters  $p(\tilde{y}, \{\theta_h, \theta_e\})$  must be taken into account. If, on the generated hyper-surface, a cut is made by a hyper-plane

877 that passes through a given vector of parameters, a so-called sampling distribution (or data distribution)  $p(\tilde{y} | \{\theta_h, \theta_e\})$  is

878 obtained, which gives the probability of the set of observations,  $\tilde{y}$ , conditioned on the values of this parameter vector. But  
 879 when the cutting of the hyper-surface is made with a hyper-plane passing through the observations of the random variable,  
 880 the concept of the likelihood function of the parameters conditioned on the observations  $\ell(\{\theta_h, \theta_e\} | \tilde{y})$  arises. The  
 881 analytical expression of the likelihood is the same as for the data distribution.



882 Although the likelihood function is not a probability function, the probability of a set of model parameters, given the  
 883 observations, is proportional to the likelihood of these parameters. In this paper, the hypothesized error model is additive and  
 884 given by:

$$885 \quad \varepsilon = y - y_s - \mu_{e|y_s} = e - \mu_{e|y_s} \quad (\text{A1})$$

886 and, when an additive error model is considered, we can write:

$$887 \quad p(y | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \ell(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \varepsilon) \quad (\text{A2})$$

888 The Conditional Probability Law establishes the following relation among the joint, marginal and conditional probabilities of  
 889 any two random variables,  $\varepsilon_1$ , and  $\varepsilon_2$ , which are conditioned to a parameter vector  $\boldsymbol{\theta}$  :

$$890 \quad p(\varepsilon_1, \varepsilon_2 | \boldsymbol{\theta}) = p(\varepsilon_1 | \boldsymbol{\theta}) p(\varepsilon_2 | \varepsilon_1, \boldsymbol{\theta}) \quad (\text{A3})$$

891 For convenience, we can split the series of random errors,  $\varepsilon = \{\varepsilon_n; n = 1 \dots N\}^T$ , into two vectors, and based on Eq. (A3) we  
 892 can write:

$$893 \quad p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = p(\varepsilon_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) p(\varepsilon_{2:N} | \varepsilon_1, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (\text{A4})$$

894 In the application of the error model defined in section 2.2, two transformations on the random errors,  $\varepsilon$ , were made to  
 895 convert them first to the studentized errors,  $\boldsymbol{\eta}$ , and finally into the standardized innovations,  $\boldsymbol{a}$ . The same transformations  
 896 must be performed on Eq. (A4) so that:

$$897 \quad p(\varepsilon_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \left| \frac{\partial \boldsymbol{\eta}_1}{\partial \varepsilon_1} \right| p(\boldsymbol{\eta}_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \sigma_{\varepsilon|q_s(t=1)}^{-1} p(\boldsymbol{\eta}_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (\text{A5})$$

$$898 \quad p(\varepsilon_{2:N} | \varepsilon_1, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \det |\mathbf{J}| p(\boldsymbol{\eta}_{2:N} | \boldsymbol{\eta}_1, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = p(\boldsymbol{\eta}_{2:N} | \boldsymbol{\eta}_1, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \left| \frac{\partial \boldsymbol{\eta}}{\partial \varepsilon} \right| = \quad (\text{A6})$$

$$p(\boldsymbol{\eta}_{2:N} | \boldsymbol{\eta}_1, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} = \prod_{t=2}^N p(\boldsymbol{\eta} | \boldsymbol{\eta}_{1:(t-1)}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} = \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} p(\boldsymbol{\eta} | \boldsymbol{\eta}_{1:(t-1)}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$$

899 In Eq. (A6),  $\det |\mathbf{J}|$  is the absolute value of the Jacobian matrix determinant for the transformation in Eq. (8). Substituting  
 900 Eqs. (A5) and (A6) into Eq. (A4) we can write:

$$901 \quad p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \sigma_{\varepsilon|y_s(t=1)}^{-1} p(\boldsymbol{\eta}_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} p(\boldsymbol{\eta} | \boldsymbol{\eta}_{1:(t-1)}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (\text{A7})$$

902 By taking into consideration Eq. (9), we can write:

$$903 \quad p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \sigma_{\varepsilon|y_s(t=1)}^{-1} p(\boldsymbol{\eta}_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} p(\boldsymbol{\eta} | \boldsymbol{\eta}_{1:(t-1)}, \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) = \quad (\text{A8})$$

$$\sigma_{\varepsilon|y_s(t=1)}^{-1} p(\boldsymbol{\eta}_1 | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \prod_{t=2}^N \sigma_{\varepsilon|y_s}^{-1} p(\boldsymbol{z} | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$$



904 Using the approximation for  $t=1$  given by Schoups and Vrugt (2010), Eq. (A8) can be rewritten as:

$$905 \quad p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \cong \prod_{t=1}^N \sigma_{\varepsilon|y_s}^{-1} p(z | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (\text{A9})$$

906 Finally, if we apply the transformation performed in Eq. (11) to Eq. (A9), we obtain the final expression of the likelihood  
 907 function:

$$908 \quad \ell(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \varepsilon) = p(\varepsilon | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \cong \prod_{t=1}^N \sigma_{\varepsilon|y_s}^{-1} \sigma_z^{-1} p(a | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\}) \quad (\text{A10})$$

909 Since we have assumed in section 2.2 that the innovations,  $\boldsymbol{a}$ , must fit a Skew Exponential Power distribution (SEP), term  
 910  $p(a | \{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\})$  is assessed by Eq. (13). To facilitate numerical computation, it is common to use the log-likelihood  
 911 function. By taking logarithms on Eq. (A10), the resulting expression is:

$$912 \quad \mathcal{L}(\{\boldsymbol{\theta}_h, \boldsymbol{\theta}_e\} | \varepsilon) \cong N \log \frac{2\sigma_\xi w_\beta}{\sigma_z (\xi + \xi^{-1})} - \sum_{t=1}^N \log \sigma_{\varepsilon|y_s} - c_\beta \sum_{t=1}^N |a_\xi|^{1+\beta} \quad (\text{A11})$$

## 913 Appendix B

914 The following lines explain in detail how to enforce the Total Laws in the inferences WLS, GL++ and GL++Bias. WLS and  
 915 GL++ implicitly fulfill the Total Expectation Law (TEL), because they are supposed to have unbiased errors. Therefore, let's  
 916 start with the Total Variance Law (TVL) which must be enforced in all cases. From the properties of the variance operator,  
 917 we can write the following expression applied to an error variance model:

$$918 \quad V[\sigma_{e|q_s}] = E[\sigma_{e|q_s}^2] - E^2[\sigma_{e|q_s}] \quad (\text{B1})$$

919 Solving for  $E[\sigma_{e|q_s}]$  in the last equation we can write:

$$920 \quad E[\sigma_{e|q_s}] = \left( E[\sigma_{e|q_s}^2] - V[\sigma_{e|q_s}] \right)^{0.5} \quad (\text{B2})$$

921 Considering the TVL given by Eq. (7) for a variance and bias error models, we can write:

$$922 \quad V[e] = E[\sigma_{e|q_s}^2] + V[\mu_{e|q_s}] \quad (\text{B3})$$

923 and, solving for the first right term and substituting in Eq. (B2) we obtain:

$$924 \quad E[\sigma_{e|q_s}] = \left( V[e] - V[\mu_{e|q_s}] - V[\sigma_{e|q_s}] \right)^{0.5} \quad (\text{B4})$$

925 In Eq. (B4) the expressions for  $E[\sigma_{e|q_s}]$ ,  $V[\mu_{e|q_s}]$  and  $V[\sigma_{e|q_s}]$  can be derived through the application of the  
 926 expectation and variance operators on Eq. (5) for the bias model, and on Eq. (6) for the variance model. This is carried out as  
 927 follows:



$$928 \quad E[\sigma_{e|q_s}] = \int_{-\infty}^{+\infty} \sigma_{e|q_s} f(q_s) dq_s = \alpha + \kappa E[q_s] \quad (B5)$$

$$929 \quad \begin{aligned} V[\mu_{e|q_s}] &= 0 && \text{if } q_s \leq q_0 \\ V[\mu_{e|q_s}] &= \int_{-\infty}^{+\infty} (\mu_{e|q_s} - E[\mu_{e|q_s}])^2 f(q_s) dq_s = \tau^2 V[q_s] && \text{if } q_s > q_0 \end{aligned} \quad (B6)$$

$$930 \quad V[\sigma_{e|q_s}] = \int_{-\infty}^{+\infty} (\sigma_{e|q_s} - E[\sigma_{e|q_s}])^2 f(q_s) dq_s = \kappa^2 V[q_s] \quad (B7)$$

931 where the expectation of the bias model is given by:

$$932 \quad \begin{aligned} E[\mu_{e|q_s}] &= \gamma && \text{if } q_s \leq q_0 \\ E[\mu_{e|q_s}] &= \gamma - \tau q_0 + \tau E[q_s] && \text{if } q_s > q_0 \end{aligned} \quad (B8)$$

933 For the hypothesis of zero-mean errors, adopted by inferences WLS and GL++, our bias model is  $\mu_{e|q_s} = 0 \forall q_s$  and then  
 934  $V[\mu_{e|q_s}] = 0$ . Therefore, substituting Eqs. (B5) and (B7) into Eq. (B4) and solving for the intercept parameter of the  
 935 variance model,  $\alpha$ , we obtain:

$$936 \quad \alpha = (V[e] - \kappa^2 V[q_s])^{0.5} - \kappa E[q_s] \quad (B9)$$

937 which is the only non-independent parameter under the zero-mean error hypothesis and corresponds to the Eq. (22).

938 Our model GL++Bias has the hypothesis of having not zero-mean errors. In this situation, we must enforce both TVL and  
 939 TEL and, as a consequence, three more error model parameters are not independent. Applying the TEL in Eq. (3) with an  
 940 error model with bias, we can write:

$$941 \quad E[e] = E[\mu_{e|q_s}] \quad (B10)$$

942 Considering our bias model defined in Eq. (5) and using the results given by Eq. (B8) we obtain:

$$943 \quad \begin{aligned} \gamma &= E[e] && \text{if } q_s \leq q_0 \\ \gamma &= E[e] + \tau q_0 - \tau E[q_s] && \text{if } q_s > q_0 \end{aligned} \quad (B11)$$

944 where,  $\gamma$  is the second non-independent parameter. First expression in Eq. (B11) corresponds to Eq. (25). If we equate both  
 945 expressions in Eq. (B11), for the streamflow threshold  $q_0$ , we can obtain the expression for parameter  $\tau$ , which is the third  
 946 non-independent parameter and corresponds to Eq. (26):

$$947 \quad \tau = \frac{E[e_1] - E[e_2]}{q_0 - E[q_{s_2}]} \quad (B12)$$



948 where, subindexes 1 and 2 distinguish the low and high streamflow populations respectively. Finally, since the  
 949 heteroscedasticity model parameters,  $\alpha$  and  $K$ , are the same for both streamflow populations, and taking into account  
 950 equations (B5), (B6) and (B7) for their substitution into the Eq. (B4):

$$\begin{aligned}
 \alpha &= \left( V[e_1] - \kappa^2 V[q_{s_1}] \right)^{0.5} - \kappa E[q_{s_1}] = \\
 951 &= \left( V[e_2] - (\tau^2 + \kappa^2) V[q_{s_2}] \right)^{0.5} - \kappa E[q_{s_2}]
 \end{aligned} \tag{B13}$$

952 From Eq. (B13),  $K$  can be solved iteratively, and it is the fourth and last non-independent parameter. When parameter  $\alpha$  is  
 953 close to zero, an approximate explicit solution is given by:

$$954 \quad \kappa = \left( \frac{V[e_1] - V[e_2] + \tau^2 V[q_{s_2}]}{E[q_{s_1}^2] - E[q_{s_2}^2]} \right)^{0.5} \quad \text{for } \alpha \approx 0 \tag{B14}$$

955 which corresponds to Eq. (24).

956

957

958 **References**

- 959 Anderson, R. L.: Distribution of the serial correlation coefficient, *Ann. Math. Stat.*, 13(1), 1–13,  
 960 doi:10.1214/aoms/1177731638, 1942.
- 961 Bates, B. C. and Campbell, E. P.: A Markov chain Monte-Carlo scheme for parameter estimation and inference in conceptual  
 962 rainfall-runoff modeling, *Water Resour. Res.*, 37(4), 937–947, doi:10.1029/2000WR900363, 2001.
- 963 Blitzstein, J. K. and Hwang, J.: *Introduction to Probability*, PRESS, CRC., 2014.
- 964 Bolstad, W. M.: *Understanding Computational Bayesian Statistics*, Wiley., 2010.
- 965 Box, G. E. P. and Tiao, G. C.: *Bayesian Inference in Statistical Analysis*, Wiley., 1992.
- 966 Box, G. E. P., Jenkins, G. M. and Reinsel, G. C.: *Time Series Analysis: Forecasting & Control.*, 1994.
- 967 Cheng, Q. B., Chen, X., Xu, C. Y., Reinhardt-Imjela, C. and Schulte, A.: Improvement and comparison of likelihood  
 968 functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme, *J. Hydrol.*,  
 969 519(PB), 2202–2214, doi:10.1016/j.jhydrol.2014.10.008, 2014.
- 970 Cowles, M. K. and Carlin, B. P.: Markov chain Monte Carlo convergence diagnostics: a comparative review, *J. Amer. Stat.*  
 971 *Assoc.*, 91(434), 883–904, doi:10.2307/2291683, 1996.
- 972 Duan, Q., Schaake, J., Andrassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L.,  
 973 Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and  
 974 Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the  
 975 second and third workshops, in *Journal of Hydrology*, vol. 320, pp. 3–17., 2006.
- 976 Evin, G., Kavetski, D., Thyer, M. and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and  
 977 autocorrelation in hydrological model calibration, *Water Resour. Res.*, 49(7), 4518–4524, doi:10.1002/wrcr.20284, 2013.
- 978 Evin, G., Thyer, M., Kavetski, D., McInerney, D. and Kuczera, G.: Comparison of joint versus postprocessor approaches for  
 979 hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resour. Res.*, 50(3),  
 980 2350–2375, doi:10.1002/2013WR014185, 2014.
- 981 Gelb, A.: *Applied Optimal Estimation*, MIT Press., 2001.
- 982 Gelman, A. and Rubin, D. B.: Inference from Iterative Simulation Using Multiple Sequences, *Stat. Sci.*, 7(4), 457–472,  
 983 doi:10.2307/2246093, 1992.
- 984 Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P. and Rieckermann, J.: Improving uncertainty estimation  
 985 in urban hydrological modeling by statistically describing bias, *Hydrol. Earth Syst. Sci.*, 17(10), 4209–4225,  
 986 doi:10.5194/hess-17-4209-2013, 2013.
- 987 Han, F. and Zheng, Y.: Multiple-response Bayesian calibration of watershed water quality models with significant input and  
 988 model structure errors, *Adv. Water Resour.*, 88, 109–123, doi:10.1016/j.advwatres.2015.12.007, 2016.
- 989 Kavetski, D., Franks, S. W. and Kuczera, G.: Confronting input uncertainty in environmental modelling, in *Science*, vol. 6,  
 990 pp. 49–68, AGU, Washington, DC., 2002.
- 991 Kavetski, D., Kuczera, G. and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory,  
 992 *Water Resour. Res.*, 42(3), n/a-n/a, doi:10.1029/2005WR004368, 2006a.
- 993 Kavetski, D., Kuczera, G. and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 2.  
 994 Application, *Water Resour. Res.*, 42(3), n/a-n/a, doi:10.1029/2005WR004376, 2006b.
- 995 Kelly, K. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, *Stoch. Hydrol. Hydraul.*, 11(1),



- 996 17–31, doi:10.1007/BF02428423, 1997.
- 997 Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, *Water Resour. Res.*,  
998 35(9), 2739–2750, doi:10.1029/1999WR900099, 1999.
- 999 Kuczera, G.: Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, *Water Resour. Res.*,  
1000 19(5), 1151–1162, doi:10.1029/WR019i005p01151, 1983.
- 1001 Kuczera, G., Kavetski, D., Franks, S. and Thyer, M.: Towards a Bayesian total error analysis of conceptual rainfall-runoff  
1002 models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, 331(1–2), 161–177,  
1003 doi:10.1016/j.jhydrol.2006.05.010, 2006.
- 1004 Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst.*  
1005 *Sci. Discuss.*, 3(4), 2145–2173, doi:10.5194/hessd-3-2145-2006, 2006.
- 1006 Laloy, E. and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS)  
1007 and high-performance computing, *Water Resour. Res.*, 48(1), 1–18, doi:10.1029/2011WR010608, 2012.
- 1008 Mantovan, P. and Todini, E.: Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J.*  
1009 *Hydrol.*, 330(1–2), 368–381, doi:10.1016/j.jhydrol.2006.04.046, 2006.
- 1010 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E.: Equation of state calculations by fast  
1011 computing machines, *J. Chem. Phys.*, 21(6), 1087–1092, doi:http://dx.doi.org/10.1063/1.1699114, 1953.
- 1012 Montanari, A., & Brath, a.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour.*  
1013 *Res.*, 40(1), 1–11, doi:10.1029/2003WR002540, 2004.
- 1014 Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, *Water*  
1015 *Resour. Res.*, 48(9), n/a–n/a, doi:10.1029/2011WR011412, 2012.
- 1016 Perrin, C., Michel, C. and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*,  
1017 279(1–4), 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.
- 1018 Pokhrel, P. and Gupta, H. V.: On the use of spatial regularization strategies to improve calibration of distributed watershed  
1019 models, *Water Resour. Res.*, 46(1), 1–17, doi:10.1029/2009WR008066, 2010.
- 1020 Reichert, P. and Mieleitner, J.: Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-  
1021 dependent parameters, *Water Resour. Res.*, 45(10), n/a–n/a, doi:10.1029/2009WR007814, 2009.
- 1022 Reichert, P. and Schuwirth, N.: Linking statistical bias description to multiobjective model calibration, *Water Resour. Res.*,  
1023 48(9), n/a–n/a, doi:10.1029/2011WR011391, 2012.
- 1024 Renard, B., Kavetski, D., Kuczera, G., Thyer, M. and Franks, S. W.: Understanding predictive uncertainty in hydrologic  
1025 modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, 46(5), W05521,  
1026 doi:10.1029/2009WR008328, 2010.
- 1027 Renard, B., Kavetski, D., Leblais, E., Thyer, M., Kuczera, G. and Franks, S. W.: Toward a reliable decomposition of  
1028 predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour.*  
1029 *Res.*, 47(11), doi:10.1029/2011WR010643, 2011.
- 1030 Sage, A. and Melsa, J.: Estimation theory with applications to communications and control, McGraw-Hill., 1971.
- 1031 Scharnagl, B., Iden, S. C., Durner, W., Vereecken, H. and Herbst, M.: Inverse modelling of in situ soil water dynamics:  
1032 accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals, *Hydrol. Earth Syst. Sci. Discuss.*,  
1033 12(2), 2155–2199, doi:10.5194/hessd-12-2155-2015, 2015.
- 1034 Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with



- 1035 correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, 46(10), 1–17, doi:10.1029/2009WR008933, 2010.
- 1036 Schoups, G., Vrugt, J. A., Fenicia, F. and Van De Giesen, N. C.: Corruption of accuracy and efficiency of Markov chain  
1037 Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, *Water Resour. Res.*,  
1038 46(10), W10530, doi:10.1029/2009WR008648, 2010.
- 1039 Smith, T., Marshall, L. and Sharma, A.: Modeling residual hydrologic errors with Bayesian inference, *J. Hydrol.*, 528(Sept-  
1040 2015), 29–37, doi:10.1016/j.jhydrol.2015.05.051, 2015.
- 1041 Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for hydrologic rainfall-runoff models:  
1042 Correlated and heteroscedastic error cases, *Water Resour. Res.*, 16(2), 430–442, doi:10.1029/WR016i002p00430, 1980.
- 1043 Sorooshian, S. and Gupta, V. K.: Automatic calibration of conceptual rainfall-runoff models: The question of parameter  
1044 observability and uniqueness, *Water Resour. Res.*, 19(1), 260–268, doi:10.1029/WR019i001p00260, 1983.
- 1045 Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W. and Srikanthan, S.: Critical evaluation of parameter  
1046 consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water  
1047 Resour. Res.*, 45(12), n/a--n/a, doi:10.1029/2008WR006825, 2009.
- 1048 Todini, E.: Hydrological catchment modelling: past, present and future, *Hydrol. Earth Syst. Sci.*, 11(1), 468–482,  
1049 doi:10.5194/hess-11-468-2007, 2007.
- 1050 Vrugt, J. a., ter Braak, C. J. F., Clark, M. P., Hyman, J. M. and Robinson, B. a.: Treatment of input uncertainty in hydrologic  
1051 modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, 44, 1–52,  
1052 doi:10.1029/2007WR006720, 2008.
- 1053 Vrugt, J. a., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. a., Hyman, J. M. and Higdon, D.: Accelerating Markov Chain  
1054 Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling, *Int. J. Nonlinear Sci.  
1055 Numer. Simul.*, 10(3), 273–290, doi:10.1515/IJNSNS.2009.10.3.273, 2009a.
- 1056 Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V. and Robinson, B. A.: Equifinality of formal (DREAM) and informal (GLUE)  
1057 Bayesian approaches in hydrologic modeling?, *Stoch. Environ. Res. Risk Assess.*, 23(7), 1011–1026, doi:10.1007/s00477-  
1058 008-0274-y, 2009b.
- 1059 Westra, S., Thyer, M., Leonard, M., Kavetski, D. and Lambert, M.: A strategy for diagnosing and interpreting hydrological  
1060 model non-stationarity, *Water Resour. Res.*, 1–24, doi:10.1002/2013WR014719.Received, 2014.
- 1061 Zheng, Y. and Han, F.: Markov Chain Monte Carlo (MCMC) uncertainty analysis for watershed water quality modeling and  
1062 management, *Stoch. Environ. Res. Risk Assess.*, 30(1), 293–308, doi:10.1007/s00477-015-1091-8, 2016.
- 1063
- 1064



1065 **Table 1. Some performance indexes for SLS and inferences with TLs, for both hydrological models.**

1066

|                              | CRR     |        |       |          | GR4J    |        |       |          |
|------------------------------|---------|--------|-------|----------|---------|--------|-------|----------|
|                              | SLS     | WLS    | GL++  | GL++Bias | SLS     | WLS    | GL++  | GL++Bias |
| <b>Log-L</b>                 | -1625.5 | -461.3 | 759.0 | 773.3    | -1819.9 | -726.1 | 742.2 | 749.9    |
| <b>NSE*</b>                  | 0.90    | 0.87   | 0.25  | 0.76     | 0.87    | 0.85   | 0.82  | 0.80     |
| <b>RMSE*</b>                 | 0.59    | 0.66   | 1.59  | 0.91     | 0.66    | 0.70   | 0.78  | 0.83     |
| <b>VE* (%)</b>               | 5.5     | 2.7    | 32.9  | 0.0      | 0.8     | -3.6   | 2.8   | 0.0      |
| <b><math>S_{\eta}</math></b> | ---     | 0.88   | 0.76  | 1.04     | ---     | 0.79   | 0.95  | 1.5      |
| <b><math>\eta</math></b>     | ---     | -0.09  | -0.73 | 0.04     | ---     | 0.07   | -0.26 | -0.28    |
| <b>Reliability</b>           | 0.79    | 0.91   | 0.50  | 0.96     | 0.81    | 0.96   | 0.81  | 0.88     |
| <b>Resolution</b>            | 3.68    | 5.66   | 4.70  | 3.94     | 3.13    | 4.45   | 4.30  | 4.40     |

\* Values for the Mean Prediction

1067

1068

1069



1070 Table 2. Maximum a posteriori (MAP) and coefficient of variation (CV) of inferred parameters, for all inferences with CRR  
 1071 hydrological model. CVs equal or larger than 0.2 are marked in red colour. Inferred parameters by the TLs enforcement are  
 1072 marked in shadow.

1073

|             |              | WLS   |             |       |             |       |             | GL++  |             |        |             | GL++ Bias |             |
|-------------|--------------|-------|-------------|-------|-------------|-------|-------------|-------|-------------|--------|-------------|-----------|-------------|
|             |              | SLS   |             | NTL   |             | TL    |             | NTL   |             | TL     |             | TL        |             |
|             |              | MAP   | CV          | MAP   | CV          | MAP   | CV          | MAP   | CV          | MAP    | CV          | MAP       | CV          |
| CRR         | <b>Imax</b>  | 10.00 | 0.06        | 2.32  | 0.18        | 2.07  | 0.17        | 0.49  | <b>0.47</b> | 0.48   | <b>0.42</b> | 0.96      | <b>0.42</b> |
|             | <b>Smax</b>  | 207   | 0.05        | 369   | 0.07        | 363   | 0.06        | 37    | <b>0.80</b> | 34     | 0.13        | 36        | 0.09        |
|             | <b>Qsmax</b> | 7.60  | 0.07        | 3.67  | 0.08        | 3.76  | 0.07        | 16    | <b>0.76</b> | 12     | 0.12        | 8.84      | 0.11        |
|             | $\alpha_e$   | 97    | <b>0.20</b> | 12    | 0.09        | 12    | 0.09        | 89    | <b>0.33</b> | 90     | <b>0.28</b> | 85        | <b>0.29</b> |
|             | $\alpha_f$   | -0.12 | <b>1.77</b> | -0.64 | <b>0.43</b> | -0.60 | <b>0.38</b> | -0.36 | <b>3.88</b> | -0.768 | <b>0.72</b> | -1.26     | <b>0.20</b> |
|             | $K_f$        | 2.46  | 0.02        | 3.14  | 0.03        | 3.15  | 0.03        | 1.98  | 0.04        | 2.05   | 0.04        | 1.83      | 0.03        |
|             | $K_s$        | 91    | 0.05        | 46    | 0.07        | 47    | 0.07        | 62    | 0.10        | 66     | 0.08        | 90        | 0.11        |
| Error Model | $\alpha$     | 0.58  | 0.06        | -0.08 | 0.08        | -0.09 | 0.17        | -0.29 | 0.09        | -0.24  | 0.05        | -0.21     | <b>0.42</b> |
|             | $\kappa$     | ---   | ---         | 0.24  | 0.03        | 0.27  | 0.02        | 0.52  | 0.06        | 0.42   | 0.03        | 0.33      | 0.09        |
|             | $\beta$      | ---   | ---         | ---   | ---         | ---   | ---         | 1.00  | 0.01        | 1.00   | 0.01        | 0.99      | 0.01        |
|             | $\xi$        | ---   | ---         | ---   | ---         | ---   | ---         | 1.03  | 0.03        | 1.04   | 0.03        | 0.95      | 0.02        |
|             | $\phi_1$     | ---   | ---         | ---   | ---         | ---   | ---         | 0.98  | 0.00        | 0.95   | 0.00        | 0.94      | 0.00        |
|             | $q_0$        | ---   | ---         | ---   | ---         | ---   | ---         | ---   | ---         | ---    | ---         | 1.96      | 0.05        |
|             | $\gamma$     | ---   | ---         | ---   | ---         | ---   | ---         | ---   | ---         | ---    | ---         | -0.25     | 0.11        |
|             | $\tau$       | ---   | ---         | ---   | ---         | ---   | ---         | ---   | ---         | ---    | ---         | -0.38     | <b>0.29</b> |

1074

1075

1076

1077



1078 **Table 3. Maximum a posteriori (MAP) and coefficient of variation (CV) of inferred parameters, for all inferences with GR4J**  
 1079 **hydrological model. CVs equal or larger than 0.2 are marked in red colour. Inferred parameters by the TLs enforcement are**  
 1080 **marked in shadow.**

1081

|                    |            | WLS   |      |        |              | GL++   |             |       |             | GL++ Bias |             |       |             |
|--------------------|------------|-------|------|--------|--------------|--------|-------------|-------|-------------|-----------|-------------|-------|-------------|
|                    |            | SLS   |      | NTL    |              | TL     |             | NTL   |             | TL        |             | TL    |             |
|                    |            | MAP   | CV   | MAP    | CV           | MAP    | CV          | MAP   | CV          | MAP       | CV          | MAP   | CV          |
| <b>GR4J</b>        | $\theta_1$ | 991   | 0.03 | 1252   | 0.03         | 1254   | 0.03        | 1847  | 0.08        | 2210      | 0.08        | 2329  | 0.08        |
|                    | $\theta_2$ | -0.71 | 0.09 | -1.36  | 0.07         | -1.37  | 0.07        | -4.92 | 0.10        | -0.55     | <b>0.21</b> | 2.89  | 0.16        |
|                    | $\theta_3$ | 153   | 0.03 | 162    | 0.04         | 163    | 0.04        | 79    | 0.04        | 81        | 0.04        | 65    | 0.05        |
|                    | $\theta_4$ | 1.55  | 0.02 | 1.66   | 0.03         | 1.67   | 0.03        | 1.54  | 0.01        | 1.67      | 0.02        | 1.58  | 0.02        |
| <b>Error Model</b> | $\alpha$   | 0.655 | 0.00 | -0.001 | <b>12.14</b> | -0.003 | <b>5.20</b> | 0.00  | <b>1.07</b> | -0.17     | 0.07        | -0.21 | 0.09        |
|                    | $\kappa$   | ---   | ---  | 0.23   | 0.06         | 0.29   | 0.04        | 0.66  | 0.12        | 0.35      | 0.04        | 0.15  | 0.14        |
|                    | $\beta$    | ---   | ---  | ---    | ---          | ---    | ---         | 1.00  | 0.01        | 1.00      | 0.01        | 1.00  | 0.01        |
|                    | $\xi$      | ---   | ---  | ---    | ---          | ---    | ---         | 0.82  | 0.02        | 0.99      | 0.03        | 0.94  | 0.02        |
|                    | $\phi_1$   | ---   | ---  | ---    | ---          | ---    | ---         | 0.97  | 0.01        | 0.93      | 0.01        | 0.93  | 0.00        |
|                    | $\rho_0$   | ---   | ---  | ---    | ---          | ---    | ---         | ---   | ---         | ---       | ---         | 1.57  | 0.13        |
|                    | $\gamma$   | ---   | ---  | ---    | ---          | ---    | ---         | ---   | ---         | ---       | ---         | -1.00 | <b>0.20</b> |
|                    | $\tau$     | ---   | ---  | ---    | ---          | ---    | ---         | ---   | ---         | ---       | ---         | -0.32 | 0.13        |

1082

1083

1084