Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-9-RC2, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Bayesian joint inference of hydrological and generalized error models with the enforcement of Total Laws" *by* Mario R. Hernández-López and Félix Francés

Anonymous Referee #2

Received and published: 21 February 2017

The study proposed by Mario R. Hernández-López and Félix Francés follows previous investigations which aim at jointly inferring the parameters of conceptual hydrological models and parameters of residual error models. This is a challenging task, since residual errors exhibit several statistical properties (heteroscedasticity, autocorrelation, asymmetry) which strongly depart from standard assumptions (iid, Gaussianity). The authors try to solve identifiability issues encountered for these parameters using constraints on the variance of these residual errors. More specifically, they decompose this variance by conditioning to the simulated streamflow.

From a theoretical point of view, this work presents a lot of interest since it focuses on global statistical properties of the residual errors and the idea of using constraints in

order to obtain some desired properties is appealing. However, the results shown in the paper do not support their conclusions (see discussion). The following paragraphs discuss numerous points partly addressed or overlooked in the manuscript.

Case study:

The French broad river catchment is a particularly wet catchment with a high annual runoff (800 mm), a high runoff coefficient (0.56), and a very small proportion of low flows. Conceptual hydrological models usually perform well for this type of catchment. As shown in Evin et al. (2013) and Evin et al. (2014), this catchment is atypical in the sense that adequate predicted streamflows (i.e. reliable and precise) are obtained even when the autocorrelation and heteroscedasticity parameters are jointly inferred. In other words, even unstable calibration schemes perform well on this catchment. I really struggle to see why the authors chose a catchment for which calibration issues are not apparent to demonstrate that their methodology solve calibration issues.

I am also puzzled by the choice of the calibration/validation period. First, they apply the hydrological models on a short five-year period (1962-1966) whereas streamflows are available for a much longer period for this catchment (until 1998). Second, they do not apply the split-sample procedure which seems essential to assess the predictive power outside the fitted period.

A major recommendation is thus to:

1. show the results of the calibration proposed in the paper on all the MOPEX catchments, as in Evin et al. (2014), 2. Apply the split-sampling procedure.

If these two requests are fulfilled, a fair comparison with the results shown in Evin et al. (2014) will be possible.

Presentation:

The current presentation of the manuscript is a bit messy. In particular, the main novelties of the paper are presented in several parts. Section 2.1. is a long introduction

to the idea of conditioning the predicted streamflow to the simulated streamflow, which was already presented at lines 62-72. Section 2.2. presents this idea in mathematical terms. Section 2.3. ("Why and when is imperative the enforcement of the total laws") tries to convince the reader that the proposed methodology is essential before the presentation of the results. I would suggest moving this section in the discussion. Finally, Section 4.4. formalizes how this idea can be applied in practice.

The general tone of the presentation gives the impression that all the previous studies for which their methodology was not applied are incorrect. In my opinion, these developments, as all the other related studies, propose calibration schemes which have different desired properties. For example, Evin et al. (2013) show that applying an AR(1) process to standardized errors usually leads to more stable results than applying the AR(1) process to raw errors. I would not say that we 'must' apply the AR(1) process to standardized errors but this approach is preferable since it leads to a more stable calibration schemes with more reliable and more precise predictive streamflows. In the proposed study, as discussed below, the obtained results are not especially impressive, and do not support statements like "The non-fulfillment of the TLs is statistically incorrect". I would appreciate if the authors could let the reader make its own opinion, without using the word 'must' too often ('must' is employed 34 times in the paper).

Interpretation of the results:

- Lines 545-546: "it is expected that the GL++ parameter estimation could be less biased than the corresponding to those classical schemes of inference". Since the 'true' parameters are unknown, a bias cannot be computed and such a statement cannot be verified. I would suggest removing this sentence.

- Lines 562-563: "In relation to the uncertainty assessment, PP-Plots in Fig. 4 show its correctness for both models." I do not understand this interpretation. In Fig. 4 (and also in Fig. 5), we clearly see that GL++ calibration scheme leads to a systematic overprediction of the streamflows, for both hydrological models.

- Lines 615-617: "Furthermore, looking at right panel of the Fig. 9, it is important to realize that the GL++Bias inference for GR4J model is the only inference that exhibits a significant contribution of parameter uncertainty to the total predictive uncertainty. This contribution seems to be underestimated in all the other performed inferences." I strongly disagree with this statement. The parameter uncertainty is related to the complexity of the calibration scheme. For example, for the SLS calibration scheme, there is only one parameter to estimate for the residual error model, which is easily identified. The parameter uncertainty is thus logically small in this case.

- Lines 717-718: "Nevertheless, the most plausible inferred value for $\theta 2$ (the closest to zero) corresponds to GL++, the most correct among these three error models." The water balance parameter $\theta 2$ in GR4J tends to compensate global under/over-estimations. In the absence of physical explanations, this parameter can thus be different from zero in order to reproduce the global volume of water. In this case, it acts as a 'bias' parameter. The fact that it is close to zero with GL++ is unclear to me, but how can the authors claim that it is the 'most correct' estimate when GL++ leads to a systematic over-estimation of the streamflows? For an unexplained reason, with GL++, $\theta 2$ is not able to compensate the excess of water produced by the GR4J model, which is certainly not a desirable feature.

Figure 15 shows that the WLS calibration scheme offers the best combination of resolution and reliability. GL++ calibration scheme leads to overpredicted streamflows and G++bias fails when the CRR hydrological model is applied (see the wide predictive limits in Fig. 9). It seems to also fail when the GR4J model is applied, since G++bias leads to meaningless parameter estimates (unrealistic estimates of $\theta 2$ in Fig. 14).

If I understand correctly the conclusion, the authors recommend the application of GL++ or GL++bias, the "fulfillment of the error model hypotheses" being the most important criteria. In my opinion, the reliability and the resolution of the predicted streamflows (second criteria) is by far the most important criteria. From an operational point of view, unreliable or imprecise predictions are useless and indicate that the calibration

scheme is inappropriate.

Parameter identifiability:

The development of more complex calibration schemes is usually difficult due to strong parameter interactions and difficulties in identifying all the parameters. This central issue has been extensively discussed in the literature (see, e.g., Renard et al., 2010) but is overlooked in the manuscript. The only exception is Figure 12, which shows that strong parameter interactions between the slope and the autocorrelation parameters are present with GL++NTL, but not with GL++. However, I suspect that other parameter interactions are present and not shown. For example, I would be curious to see the correlations between $\theta 2$ and the other parameters with the GL++ calibration scheme, in particular with the parameter of the bias model in Eq. (5). That would explain the high values of $\theta 2$ in Fig. 14.

At lines 639-641, the authors claim that "This incorrectness generates problems, mainly related with spurious parameter interactions, affecting the inference results and making them unsuitable and possibly non-robust (Evin et al., 2014). This section will demonstrate that not enforcing the TLs is, at least, one of the most important causes of these problems." To be demonstrated, the authors must show that these parameter interactions are systematically removed, and not only for a couple of parameters.

Bias:

Since hydrological models usually have parameters affecting the water balance (as the $\theta 2$ parameter of the GR4J model), I struggle to see how the parameters of the hydrological models can be jointly inferred with the parameters of a bias model. I suspect that strong parameter interactions lead to the meaningless estimates of $\theta 2$ with GL++bias in Fig. 14. Furthermore, in Table 3, the MAP value of the γ parameter is exactly -1. This rounded value could indicate that a lower bound has been set to this parameter and that it cannot be identified with GL++bias.

Kurtosis:

The β parameter indicates the kurtosis of the SEP distribution. It is associated to the forth moment and can be interpreted in terms of flatness of the distribution. In this study, as in Schoups and Vrugt (2010), this parameter always hits the lower bound (β=1) and seems difficult to identify. I would suggest trying alternative calibration schemes without the kurtosis component. Calibration schemes with a Gaussian distribution instead of a SEP distribution in GL++ (without the TLs constraints) corresponds to calibration schemes tested in Evin et al. (2014) and could be interesting to compare to the calibration schemes of the manuscript.

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

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-9, 2017.