We thank the reviewer for nicely summarizing the key aspects of our study and pointing out their importance. We have addressed all of his comments point by point and we have tried to respond as short and clear as possible. Moreover, we enormous appreciate the interest shown by reviewer, even in the specific details of our work as figures, tables, etc. We have revised the manuscript based on the suggestions and advice of the reviewer and we have done a big effort in the next replies. We hope that our answers successfully address his concerns and requirements, and the proposed modifications for the revised manuscript will be accepted for publication.

Reviewer's comment #1

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 solves calibration issues.

Reply

Reviewer is partially right. Previous papers of Evin et al. (2013, 2014) did not show any kind of problems with the FB basin. Reviewer has to realize that Evin et al. consider a Gaussian distribution for the innovations, in all cases. However, a posteriori checks that they perform on their inferences show that Gaussian hypothesis is non-realistic, even for the FB basin. In order to improve this point, our manuscript considers a SEP distribution, which allows the additional possibility of modeling the skewness and the excess of kurtosis for the innovations. Having a more flexible inference (two more parameters) is a key difference: although it may seem insignificant, it is the cause which produces the known inference problems (enlargement of the uncertainty bands, getting spurious parameter interactions and yielding unidentifiable autocorrelation parameter), also on FB basin, except when TLs are considered.

We would suggest to the reviewer the reading of the reply **RC1#5** to the reviewer-1.

Reviewer's comment #2

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.

Reply

A 5-years calibration period is exactly the same period (in length and range of dates) used in Schoups and Vrugt (2010): same basin, same period, and same CRR model (additionally, we have also used GR4J). About this matter, Schoups and Vrugt (2010) state: "*Experience suggests that 5 years of daily streamflow data contains enough information about the parameters of conceptual rainfall-runoff models…*". Brigode et al. (2013) states that "*there is no clear consensus on the minimum length of calibration period for rainfall-runoff models, which is probably attributable to the specificity of the catchments and models used in those studies. Specifically … Anctil et al. (2004) obtained good GR4J performance with 3 to 5 year calibration periods and Perrin et al. (2007) showed that the calibration of the GR4J … with the equivalent of only 1 year of data can provide acceptable performance*". Actually, in our experience with better models (an also with more parameter involved), calibration period is possible with just 1 year of daily data (Frances et al., 2007).

We will add this information in the revised manuscript.

Regarding the issue of the validation using a split-sample procedure, the topic of our paper is mainly about the hydrological model calibration via a statistical methodology. The paper tries to explain and solve the theoretical pitfall which is committed in a joint inference when TLs are not considered. Therefore, it is only necessary the comparison of the methodology with and without the mistake during calibration. We do not intend to demonstrate if GL++ and GL++Bias error models and the used hydrological models are absolutely the best ones for the analyzed case study: we agree, this would require a thorough validation at different periods to the calibration one. But, this is not the goal of the manuscript and we wanted to be as short as possible.

We will make this clarification in the goals defined in the introduction of the revised manuscript.

Reviewer's comment #3

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.

Reply

Regarding the structure of the document, we will change it according to the suggestions of reviewer-1 defined in RC1#9.

The proposed methodology tries to transpose, what occurs with the error marginal and conditional pdf's in a SLS inference, to other more "sophisticated" schemes of inference. We have not coined the Total Laws. We have only understood that these Laws always must (from a theoretical point of view) be fulfilled: by its own (as in SLS) or enforcing them. For this reason, we do not have to wait for results to assert that Total Laws must be theoretically always fulfilled. For this reason, the manuscript introduces the theoretical concept of TLs from the beginning, without waiting for the case study results.

<u>Reviewer's comment #4</u>

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).

Reply

We are very sorry about the impression that the manuscript produces on the reviewer. Our paper brings to light theoretical concepts not used in previous researches. This does not mean that previous papers are useless. Surely our research will be also overcome in the future, but this is not the important thing. Our deep and sincere intention is only making a small but useful contribution to the hydrological community.

Having said that, regarding the sentence "*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…*". If reviewer reads the comment **RC1#5**, he will have the arguments to understand the reason why we say "must" in that sentence: from our experience, recommendation of Evin et al. (2013) is necessary, jointly with the application of TLs, to avoid the "feared" and meaningless uncertainty band enlargement, even in the FB basin. Anyway, thanks to the comment of reviewer, we become aware that we have abused of word "must". We will re-read the manuscript for correcting this issue. For now, **Line 398** of original manuscript will be modified as: "*Firstly we have included the recommendation given by Evin et al. (2013). As they mentioned* *and our experience confirms, errors should be studentized before applying an autoregressive model on them.*"

Regarding the sentence "*I would appreciate if the authors could let the reader make its own opinion, without using the word 'must' too often*". We are very sorry for this abuse of "must". Some of them are related with our belief that TLs always must meet, **from a theoretical point of view**. We have re-read the entire manuscript in order to improve and to smooth the expressions in the revised manuscript. For example, a compromise solution will be using:"must, from a theoretical point of view" and "should" in all other cases.

Reviewer's comment #5

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.

Reply

We take and share that idea from Schoups and Vrugt, (2010) (among others) and the proper reference is included in our manuscript in the same **Line 546**. They state: "***Violation of SLS assumptions*** *may introduce bias in estimated parameter values and affect parameter and predictive uncertainty [Thyer et al.,2009]*".

They also state: "***Robustness of the GL inference results*** *can be attributed to three factors: (1) by accounting for heteroscedasticity less weight is given to high flows, making the inference less sensitive to large flow events in different calibration data sets; (2) long tails of the Laplace distribution allow for a larger number of large errors, which again induces robustness against outliers and random variations in large flow events; (3) accounting for autocorrelation in the residual errors filters out measurement, model input, and model structural errors, resulting in less biased and more consistent parameter estimates [Vrugt et al., 2005]"*

From those three references (Schoups and Vrugt, (2010); *Vrugt et al., (2005); Thyer et al., (2009)*) and others as Sorooshian and Dracup (1980) or Sorooshian and Gupta (1983), and from our own experience, it can be concluded that: a correct error model which correctly considers those three elements (heteroscedasticity, dependence, non-Gaussianity) should theoretically yield more robust parameters (with less bias) and more robust predictive distributions, than the classical error models.

As we mention in **Lines 547-549**, "…*this is the reason for the poor performance shown by the biased prediction of the hydrological model: the most plausible parameter set, for both hydrological and error models, brings out (in form of a prediction bias) the deficiencies in the hydrological model and/or in the input data*". That is to say, when there are problems with data and/or with the hydrological model conceptualization, there is a tradeoff between getting a good fitting of the hydrological model to the observations and getting a reliable calibrated parameter set. To get both things is theoretically impossible. This fact is shown by the GL++ error model of our manuscript.

In the revised manuscript, we will add to this paragraph, all mentioned references. Most of them were already included in other parts of the original manuscript. We will also complete the explanations.

Reviewer's comment #6

- 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.

Reply

PP-Plots are a useful tool which is able to show us two different but related aspects about the predictive distribution. The first one is related with the bias of the predictive distribution, that is to say, with the shifting of the predictive distribution relative to the observations. Only when the PP-Plot crosses the diagonal line at probability 0.5, the predictive distribution is unbiased; otherwise we will have a systematic Overprediction/Underprediction. The second aspect shown by PP-Plots is about the correctness of the width of the bands, that is to say, the "quality" of the uncertainty estimation. For example, an inverse S-shaped PP-Plot (typically related with SLS error model, as reviewer can see looking at blue solid lines in **Figure 4)** indicates an Overestimation of the predictive uncertainty. On the contrary an S-shaped PP-Plot would indicate an Underestimation of the predictive uncertainty. Reviewer has to realize that these explanations are valid if the axes configuration is as in our manuscript. Some papers have inverted the axes; therefore the explanations will be also inverted. In the manuscript, **Line 487**, we reference the papers in which reviewer can find a much better exposition about this tool.

Therefore, concerning the **Lines 562-563** about GL++, the explanation is as follows. The first aspect to notice is the related with the bias of the predictive distribution: since the black solid line in **Figure 4** does not cross to the diagonal at 0.5 (actually in any point) the predictive distribution is clearly biased. As the PP-Plot is under the diagonal there is an Overprediction: the more the distance of the PP-Plot to the diagonal, the greater the bias of the predictive distribution. Regarding the second aspect, the uncertainty assessment which is related with the amplitude of the bands, as the PP-Plot does not have an S-shape, we can say that the uncertainty estimation is correct.

We will improve these explanations in the revised manuscript.

Reviewer's comment #7

- 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.

Reply

In general, it is true that with less number of parameters, the uncertainty is smaller, if the model is correct. We found underestimation of parameter uncertainty using SLS as it was also found by Schoups and Vrugt (2010) and Thyer et al. [2009] among others. This occurs mainly due to fact that SLS neglects all sources of uncertainty different to measurement uncertainty. In hydrological modeling this can lead to an incorrect parameter inference.

What we say in **Lines 615-617** is just a description of the results. GL++Bias is the only inference which exhibits wide yellow bands (predictive uncertainty due to parameters). And the responsible of this widening is the θ2 parameter of GR4J model. In **Table 3**, the coefficient of variation for θ2 in GL++ is larger (0.21) than in GL++Bias (0.16). That is to say, θ2 parameter is more uncertain in GL++ than in GL++Bias. But, we can observe in **Figure 7–right**, how in GL++, parameter θ2 is not able to propagate its larger uncertainty through the hydrological model: yellow band is hardly appreciable. However in GL++Bias, θ2 smaller uncertainty produces a wider yellow band. The reason is that, in GL++Bias, the hydrological model is more sensitive to the value of θ2 than in GL++. In other words, in GL++Bias, the data support strongly the optimal (MAP) value adopted by θ2, since variations from this optimal are not possible without modifying considerably the output.

The aim of these explanations in the manuscript is merely to make a description of what occurs, and the main comparison is between GL++ and GL++Bias, since both are complex error models. The following **comments #8 to #11** perhaps clarify better the issues about θ2 parameter.

Of course, we will add these explanations in the revised manuscript.

Reviewer's comment #8

- Lines 717-718: "Nevertheless, the most plausible inferred value for θ2 (the closest to zero) corresponds to GL++, the most correct among these three error models." The water balance parameter θ2 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++, θ2 is not able to compensate the excess of water produced by the GR4J model, which is certainly not a desirable feature.

Reply

Firstly, θ2 is a hydrological parameter whose definition is groundwater exchange coefficient for the deep (or regional) aquifer of the basin. As reviewer correctly underlines, θ2 is able to compensate (and close) the global water balance. But, for doing that, a deep aquifer should exist in the basin. If this is not the case, θ2 should not compensate anything. Therefore, since any deep aquifer exists in this basin (**Lines 711-712**), θ2 parameter should not do anything except adopting a zero or near-zero value. If we want a parameter which acts "as a bias parameter" we should consider it, but in an explicit way in the error model, as GL++Bias does.

Reviewer is right: GL++ leads to a systematic over-estimation of the streamflows. GL++ is acting as a (very general) diagnostic tool. GL++ says to us: "these are the best results that your model and your data are able to yield; if you want better results improve the data, the model or both."

As we previously mentioned (**reply #5**), GL++ brings to light the tradeoff between getting a good fitting of the hydrological model to the observations and getting a reliable calibrated parameter set. To get both things is impossible when model and/or data have problems.

Reviewer's comment #9

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 θ2 in Fig. 14).

Reply

Regarding the WLS error model, reviewer can see our replies to reviewer-1 **RC1#3** and **RC1#6**.

We agree, GL++ calibration scheme leads to over-predicted streamflows. In fact as previously mentioned, GL++ is acting as an overall diagnostic tool on the inference: GR4J model plus the used input data are not able to reproduce the observations, as SLS makes us believe.

GL++Bias with CRR does not fail. **Figure 9** shows an "ugly" uncertainty band, but its reliability is quasi-perfect, as the green solid line in **Figure 4** shows. Given that the check of the other error hypotheses is at least acceptable (as **Figure 8** shows), we can conclude that the problem is in CRR model and not in the estimated uncertainty band: CRR yields results with large uncertainty in our case study. For finding a faulty inference, reviewer can see **Figures 7 and 13**, which compare GL++ error model with and without the TLs enforcement, for the two hydrological models: we can affirm that GL++NTL (**Figure 13**) fails.

We agree with reviewer: GL++Bias error model is not the best error model for GR4J, and probably it is not also for CRR: it was not our goal to find the best error model for the analyzed case study. However, by checking the error hypotheses fulfillment we can affirm that they are far better than SLS and WLS (see **section 5** of original manuscript). Regarding the θ2 parameter, we mention in **Lines 717-718** that the most correct θ2 estimation is the closest to zero (yielded by GL++). As GL++ is showing us that GR4J overpredicts (it needs dewater the basin), therefore a correct bias model would be able to correct this situation. Besides, looking at **Figure 14**, reviewer can see how the application of the bias model, at least moves θ2 parameter from the negative value in GL++ to a positive one. Therefore, it seems plausible that a better bias model be able to move the θ2 value towards zero. But, this is out of the scope of this paper and it will be done in future research.

<u>Reviewer's comment #10</u>

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.

Reply

As we have mentioned previously, the aim of the paper is not to find "the best" error model. However, GL++ and GL++Bias are far better than classical error models, as we exposed previously. The only thing that we recommend is the enforcement of the Total Laws on inferences which consider jointly the variance and dependence models for the errors. The error models for the variance, the bias and the autocorrelation that we have used are susceptible of being improved, without any kind of doubts: but this is not the aim of the manuscript.

Regarding the issue of how important is the reliability and resolution of the predictive distribution; we want to make an instructive reflection. Having a good reliability or resolution should not be important when the errors exhibit high autocorrelation, which indicates a misspecification of the hydrological model: e.g. this occurs with WLS error model. It is important to remember that, model misspecifications diminish the predictive power of the misspecified models, although apparently, they can exhibit good performance in calibration. The underlying cause of this good performance in calibration is the over-fitting through a "forced value" for parameters, rather than the correctness of the modeling. Kirchner, (2006) states "*[…] to advance the science of hydrology, as opposed to the operational practice of hydrology (that is, to improve our understanding of how hydrologic systems work), we need to know whether we are getting the right answers for the right reasons […] advancing hydrologic science, rather than providing better predictions for operational purposes, although of course one hopes that the former may lead to the latter*".

Therefore, from a **"model calibration point of view"**: if we strive for the error modeling (using error models with an adequate complexity) as we do it for the hydrological modeling, we will obtain benefits as, more reliable hydrological parameter estimation and also more robust predictive distributions. Therefore, we claim: First: The benefits of using a complex error model only arise by performing a joint inference of hydrological and error model parameters. Second: using a complex error model, with the aim of exploiting its benefits, requires obviously the fulfillment of its hypotheses by the inferred errors. Third: The definition of the error model through the definition of its conditional distributions, as made in our paper and all related ones, requires the check of consistency between the error marginal and these error conditional distributions. When this consistency does not occur by its own, we propose to use the TLs enforcement.

That said, if the priority is related only with the **"operational point of view"**, rather than with trying to make a reliable parameter inference jointly with getting reliable predictive uncertainty estimations, the problem is very different. There also exist other much more efficient methodologies, generally used for this tasks (papers from authors Krzysztofowicz and Todini are good references). These methodologies are based on calibrated models (including black-box models), with any SLS-related calibration method, which feed an uncertainty Post-Processor. We explain this in **Lines 98-132** of the manuscript. Reviewer can also read our reply **RC1#6** to reviewer-1 for more ideas about the relative importance of reliability and resolution of the predictive distribution, in a model calibration context.

<u>Reviewer's comment #11</u>

Parameter identifiability:

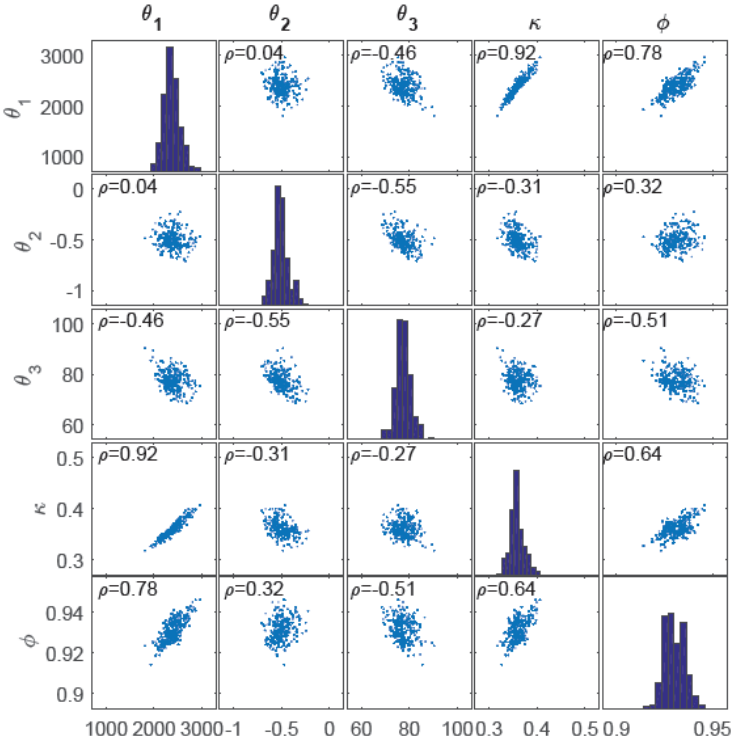
The development of more complex calibration schemes is usually difficult due to strong parameter interactions and difficulties in identifying all the parameters. This central issue has been extensively discussed in the literature (see, e.g., Renard et al., 2010) but is overlooked in the manuscript. The 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 θ2 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 θ2 in Fig. 14.

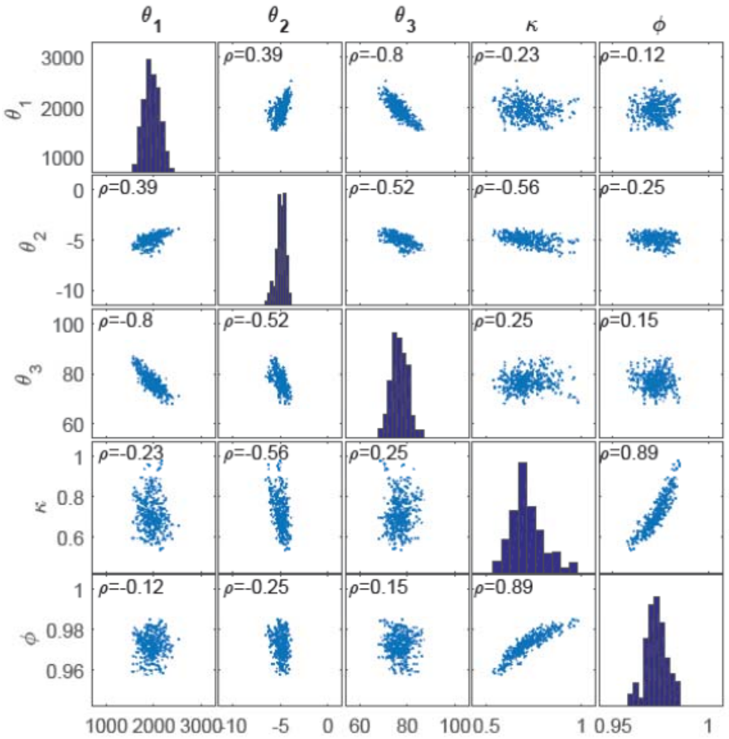
Reply

Right: this is a very important topic even for us! Unfortunately, with the aim of not lengthening the article too much we only presented the result and figure more representative of the main objective of the paper. **Figure 12**, because this figure represents a faulty inference of parameter phi, but only when TLs are neglected. As we explain in **RC1#5**, "*Besides (and the most important), it can be observed the extremely high inferred MAP value for phi (about 0.99) when TLs are neglected. In this case the posterior distribution of phi shows extreme asymmetry, with the mode at the value of one, the upper bound value for autocorrelation parameter. This problem was also reported in Scharnagl et al. (2015) for their Likelihood2. From our point of view, this is a synonym of having a non-identifiable distribution for the autocorrelation parameter,* *since for phi→1, the AR(1) process becomes nonstationary, as explained in Box et al. (1994).”* Other interactions among parameters could be possible, but they have nothing to do with instability of the inference.

Anyway, we will try to calm the curiosity of reviewer by showing three interesting figures about inferences with the GR4J model (not shown in original manuscript). The two first figures are about the most significant parameter interactions on GL++ and GL++NTL inferences. Top figure is from GL++ and the bottom one from GL++NTL (that is to say, GL++ without TLs). The main points to remark on top figure are:

1- GL++ does not show high correlation among hydrological parameters
2- The highest correlations are between θ1-kappa and θ1-phi. However, all parameters are perfectly identifiable, and this inference does not suffer the enlargement of the uncertainty bands.
3- Correlations of θ2 are moderate or low
4- Correlation between kappa-phi is moderate (0.64)



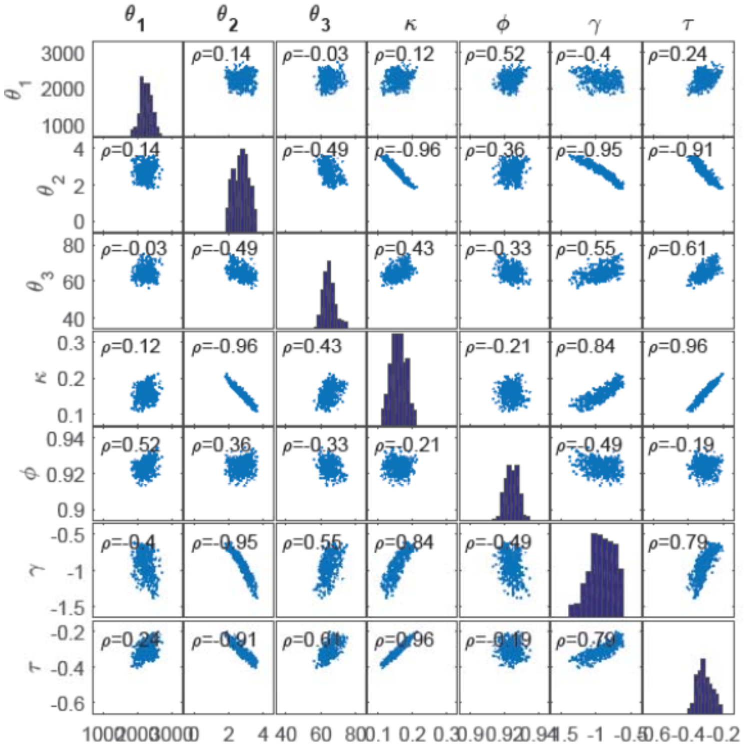


The main points to remark on bottom figure, about GL++NTL are:

1- GL++NTL shows a significant increment in correlations among hydrological parameters θ1-θ2 (0.39 vs 0.04) and θ1-θ3 (0.80 vs 0.46). Why this undesirable increment in correlations between hydrological parameters? They should not to be so strongly correlated, as GL++ shows. These increments in the hydrological parameter covariance matrix (Jakeman and Hornberger ,1993), when TLs are neglected, indicate that hydrological parameters catch less information from data than they do when inference considers TLs. Hence, from this point of view, GL++NTL inference is less efficient than GL++. This seems to be another more symptom of the theoretical pitfall.
2- Correlations of θ2 are moderate but larger than in GL++, for θ1-θ2 and θ2-kappa
3- θ2 MAP value deserves special attention, because it takes the highest absolute value for all inferences: θ2=-4.92. If we accept that θ2=0 should be the correct value, NTL inference yields the farthest value from the correct, even farther than SLS (θ2=-0.71) or WLS (θ2=-1.37).
4- Correlation kappa-phi is the highest (0.89). However, phi is identifiable with GR4J, differently to what occurs with CRR model (see **Figure 12** in original manuscript).

Therefore, with GR4J there are several cases of parameter interactions. The most harmful, from our point of view, are among hydrological parameters: but these occur only in GL++NTL inference. With CRR we found also the problem of a parameter phi non-identifiable (**Figure 12** in original manuscript).

Finally, we will show the figure of correlations among parameters, corresponding to GL++Bias (also with GR4J).



The main points to remark on figure, about GL++Bias are:

1- Hydrological parameters do not exhibit high correlations (as in GL++)
2- All parameters are identifiable, including “gamma” for the bias model (this replies to the **comment #13**)
3- Reviewer is very right. The highest correlations in matrix correspond to those between θ2 parameter and variance and bias error parameters: θ2-kappa (0.96), θ2-gamma (0.95) and θ2-tau (0.91).
   But, where is the problem? We previously explained (see **reply #8**) that the aim of the bias model, in our case study, is to explicitly recognize the necessity of dewatering the basin. We also mentioned that θ2 is not the parameter which has to do this task. Therefore the error model is complying with its function of trying to substitute to the θ2 parameter in the dewatering task. They are so correlated because they are doing the same function: trying to manage the water excess in the basin. But, in our case, the (erroneous) bias model is strongly dewatering the basin; therefore θ2 becomes positive (to bringing water from… some “virtual” aquifer) in order to compensate the excess of drainage by bias model. In short: 1- a correct bias model is necessary to allow θ2 gets a zero-value; 2- our bias model is not the proper one.

For this comment and the two next, we will add these figures and comments as annex in the revised manuscript, if editor allow us, with the corresponding condensed paragraphs in the main text.

Reviewer's comment #12

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.

Reply

With previous reply to **comment #11** for GR4J and with **Figure 12** for CRR, we demonstrate how the actual problems with parameter interactions only arise in the inferences without Total Laws. The main detected problems are: 1- With GR4J the spurious interactions among hydrological parameters and 2- With CRR, the non-identifiable autocorrelation parameter. We are not able to detect, in our case study, another more important effects related with parameter interactions. These problems appear when TLs are neglected and they do not when we enforce TLs. But see our long previous reply!

Reviewer's comment #13

Bias:

Since hydrological models usually have parameters affecting the water balance (as the θ2 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 θ2 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.

Reply

See our previous reply to **comment #11**

Reviewer's comment #14

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 (sic) 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.

Reply

We consider the SEP distribution of Schoups and Vrugt (2010), because is more general than Gaussian, but if the inference it requires, would also allow Gaussianity; therefore we have two more error parameters (skewness and kurtosis). As it is shown in Evin et al. (2013, 2014) results, these two additional statistical properties are necessary to model errors more correctly.

We disagree with the reviewer's affirmation "*this parameter always hits the higher bound (β=1) and seems difficult to identify*" The value of "1" is a valid and identifiable one, since it is which yields the maximum kurtosis. Moreover, our GL++Bias inference exhibits (see **Figure 8**) an excess of kurtosis that even SEP is not able to reproduce. Perhaps the Skewed Student distribution, as the used by Scharnagl et al. (2015) would have a better behavior than SEP.

Of course, it would be interesting making other comparisons as with the Skewed Student distribution, using more hydrological models, experiments on more basins and also improving the error variance and dependence models. But we consider it is out of the scope of this manuscript and would enlarge unnecessarily the paper. More research is on-going for future papers.

References

Anctil, F., Perrin, C., Andréassian, V., 2004. Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. Environ. Model. Software 19 (4), 357–368. http:// dx.doi.org/10.1016/S1364-8152(03)00135-X.

Brigode, P., Oudin, L. and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, J. Hydrol., 476, 410–425, doi:10.1016/j.jhydrol.2012.11.012, 2013.

Frances, F., Velez, J. I. and Velez, J. J.: Split-parameter structure for the automatic calibration of distributed hydrological models, J. Hydrol., 332(1–2), 226–240, doi:10.1016/j.jhydrol.2006.06.032, 2007.

Jakeman, A. J. and Hornberger, G. M.: How Much Complexity Is Warranted in a Rainfall-Runoff Model ? are good predictors of streamflow and, Water Resour. Res., 29(8), 2637–2649, doi:10.1029/93WR00877, 1993.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42(3), n/a--n/a, doi:10.1029/2005WR004362, 2006.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., Mathevet, T., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. Hydrol. Sci. J. 52 (1), 131. http://dx.doi.org/10.1623/ hysj.52.1.131.

Vrugt, J. A., B. A. Robinson, and V. V. Vesselinov (2005), Improved inverse modeling of flow and transport in subsurface media: Combined parameter and state estimation, Geophys. Res. Lett., 32, L18408, doi:10.1029/2005GL023940