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 and we have tried to respond as short and clear as possible. We also indicate the changes we will make in the revised manuscript by considering reviewer's suggestions. We hope that our answers and proposed modifications, for the revised manuscript, will be considered and accepted.

Reviewer's comment #1

The manuscript investigates residual error models for the calibration of conceptual hydrological models. The authors rightfully point out that the previous literature has identified failures in the joint calibration of hydrological and error model parameters under particular conditions, e.g. for the GL approach (Schoups and Vrugt 2010) and for the WLS-AR1 approach (Evin et al 2013, 2014). I generally found the study interesting as it explored an important aspect of hydrological model calibration using statistical techniques. The implications of the Total Variance Laws on uncertainty estimation in hydrological models is certainly worthy of research attention and this manuscript does provide some insights towards that aim. There are other interesting results, e.g., some of the analyses around Fig 15 are instructive and visually well presented.

Reply

We totally agree with these comments. Moreover, research on the application of Total Laws in the joint inference with Generalized Error Models is not only "interesting" but it is necessary, if we want to perform a statistically correct full joint inference. The main idea is that a predicted state variable and its error form a joint pdf (this also occurs in SLS!). In our case this pdf is $p(e,q_s)$. In this joint pdf (as in any bivariate pdf) Total Laws must be fulfilled. This is the cornerstone of the paper.

In a general case (SLS is the exception), Total Laws are not fulfilled if they are not explicitly enforced. Neglecting these Laws is a statistical pitfall, independently of any other consideration. In other words, the main problem can be noticed from a theoretical level, without the need of more empirical results. However, these empirical experiments are useful to prove that not committing the theoretical mistake produces the disappearance of the identified failures detected in literature with more "sophisticated" error models than SLS.

Reviewer's comment #2

1. METHOD SELECTION

The aim of the study is to address the failure - via unstable/explosive prediction limits - of joint inference approaches. However, this failure has been reported only specifically when inferring the autocorrelation parameter rho - the manuscript notes this on lines 39-40 citing Evin et al 2014, but maybe overlooks that Evin 2014 show that if rho is fixed, there is no instability.

Reply

From our point of view, the failure is the non-fulfillment of the Total Laws rather than the more or less "explosivity" in the enlargement of the prediction limits. This enlargement (explosive or not), as well as other spurious effects, is a mere consequence of the main failure.

The paper is about “Joint Inference”. But, if we fix any parameter, either from the error model or from the hydrological model, then we are not doing a full joint inference. Fixed parameters do not interact with the other during the inference process. Therefore, if rho is fixed, it will not interact with anything. Fixing parameters is not an option for us, since this way to proceed can produce biased parameters and incorrect predictive distributions as it is explained in **lines 49-55** of the original manuscript.

In **lines 38-40** we say “*They* (Evin et al.2014) *concluded that the joint inference could be non-robust due to multiway interactions between the hydrological parameters (related with the water balance in their case study) and the error model parameters, particularly, heteroscedasticity and autocorrelation error model parameters*”. In the revised manuscript we will add a clarification in the introduction, which it could be as follows: *“The cause of this “non-robustness” is not the possible interaction among parameters. The cause is a statistical one: the non-fulfillment of Total Laws. All other detected problematic (or not) effects, as the possible interactions, are a consequence of this severe statistical mistake.”*

Reviewer’s comment #3

The manuscript then uses the SLS and WLS approaches as a major part of the analysis - even though neither of these methods have an autocorrelation parameter, let alone infer it! So WLS (and SLS), even if they are joint inference methods, do NOT suffer from the instability the study is trying to resolve, and this has already been known from the cited previous studies.

Reply

Reviewer is right. It is obvious SLS and WLS do not suffer instabilities in the inference process. This occurs because these methods do not infer simultaneously the structures of variance and dependence of the errors. In **lines 658-660** paper says: “*this research demonstrates that when the error model considers both the autocorrelation (AR) and the heteroscedasticity models, not enforcing the TLs has a significant effect on the result of the inference*”. In Conclusions, **lines 813-815**, this idea is repeated. Therefore we clearly specify in which inferences we have detected and solved the problem. Not in SLS. Not in WLS. Only in those inferences which are similar to GL++ and GL++Bias. As it is shown in **Figure 1-a**, SLS always fulfills TLs without explicitly enforcing them. However, a priori WLS does not: as it will be explained later, WLS method yields the same results either with TLs fulfillment or without it.

Analysis of SLS or WLS are included as a reference to compare two classical error models, which generally are not valid for inferences in hydrological modeling, with other two new error models GL++ and GL++Bias, which try to be more appropriate to the errors generated in hydrological modeling.

Finally, we do not agree with the affirmation “*The manuscript then uses the SLS and WLS approaches as a major part of the analysis*”. Objectively, we have dedicated 77 lines to SLS and WLS, and 96 lines to GL++ and GL++Bias, within the Results section. In other sections this difference is considerably larger.

Reviewer's comment #4

To demonstrate how the TVL approach removes the instability shown by Evin et al 2014, the WLS-AR1 approach from Evin et al should be clearly included in the analysis, which is the error model where the instability actually occurs.

Reply

We agree and we have done the experiment: our GL++NTL model is the equivalent approach to the WLS-AR1. It is not exactly the same because we do not assume the Gaussian distribution for the innovations as it is made in Evin et al. (2013, 2014). We consider the SEP distribution of Schoups and Vrugt (2010), which is more general than Gaussian; therefore we have two more error parameters (skewness and kurtosis). As it is shown in Evin et al. (2013, 2014) results, considering these two additional (parameters) statistical properties is necessary for trying to model the errors more correctly. On the other hand, GL++ would be the WLS-AR1 approach, plus the enforcement of Total Variance Law.

In the revised manuscript, we will clarify better this equivalence which should be in the present section 4.3.

Reviewer's comment #5

2. CASE STUDY CATCHMENT

The catchment used in the manuscript to demonstrate its contribution is the French Broad River from the MOPEX database. This is a very strange choice of catchment for the given research objectives, because it is one of relatively few catchments where pretty much every residual error model has performed well, including the original GL approach of Schoups and Vrugt (2010), and the joint WLS-AR1 scheme of Evin et al 2013, 2014, as can be seen from those papers. In this respect it should be clear that using such a case study catchment cannot provide supporting evidence of the conclusions. If the authors wish to demonstrate they have "solved" the problems with the above error models, I think it should be obvious that the case study should *at least* use a catchment where the old models fail in the sense of producing clearly explosive prediction limits (the problem the study is trying to solve) and the new model shows significant improvement (the outcome the study is trying to achieve).

3. GENERAL SUPPORT FOR CONCLUSIONS

I struggle to see empirical evidence in support of the conclusions, which begin with "This paper has addressed the challenging problem of jointly estimate hydrological and error model parameters in a Bayesian framework, trying to solve some of the problems found in previous related researches, as in the second case study of Schoups and Vrugt (2010) as well as in Evin et al. (2014), among others".

Reply

We do not agree at all with the affirmation "*… using such a case study catchment cannot provide supporting evidence of the conclusions*". It is just the opposite! This basin has been used in other papers similar to ours. The "kindness" shown by FB basin when is modeled, verified in all these papers, was the reason which led us to choose it. We did not want a complicated basin (actually either a complicated hydrological model), to avoid introducing more difficulties into the main problem addressed by our paper: The Total Laws fulfillment. We thought that showing problems in the inference

with this “easy” basin would be more instructive than using a difficult basin which yields a bad performance, even with a classical inference method as SLS.

Reviewer (and readers) has to realize that our experiments consider a full error model with: varying variance, internal dependence, skewness, kurtosis and even a varying bias model in the full case (GL++Bias). From our knowledge, this has not been made before with success. The more similar research to ours is the paper of Scharnagl et al. (2015). They tried (without success) the joint inference with a skewed Student distribution for the innovations (including both skewness and kurtosis parameters) and with variance and dependence error models which are similar to ours.

Schoups and Vrugt (2010) did not consider a so thorough error model: in their first case study with FB basin, they did not infer the skewness parameter jointly with all the other. We reproduced the same first case study in Schoups and Vrugt (2010), and we did also the experiment including the skewness parameter in the joint inference (not included in our manuscript). In the first case, there were not problems, but in the second one, it was not possible to find a proper inference: after 1.5E6 MCMC iterations, the Gelman-Rubin criterion of convergence is not yet reached, whereas if skewness is neglected, convergence is got with only 0.1E6 MCMC iterations.

It is important to remember that Schoups and Vrugt (2010) modeled the variance of the innovations, instead of modeling the variance of the errors as it was proposed in Evin et al. (2013). However, Total Laws must be also fulfilled on innovations: their marginal and conditional distributions are also related! As an experiment, we also tried (not included in our manuscript) the same first case study with TLs enforcement on innovations (instead of on errors); that is to say, the same error model of Schoups and Vrugt (2010), but including skewness and applying the Total Variance Law (TVL): convergence was reached at only 0.1E6 MCMC iterations.

As reviewer points out, Schoups and Vrugt (2010) did not find any problem with FB. Neither the methodological problem solved in Evin et al (2013) nor the problem with inferring also the skewness parameter. But these problems were there, although they were not visible, as we explain in the following: the inference with skewness parameter and TVL enforcement showed us the known and “feared” enlarged uncertainty band: note that the variance model is on innovations! However, if we apply the recommendation of Evin et al. (2013) the enlargement disappears, which is our case with CRR and GL++. As summary, the first case study of Schoups and Vrugt (2010), with FB basin, can show the detected problems in joint inference, with the following modifications:

1- Including skewness in the joint inference: the MCMC problem does not converge
2- Including skewness and enforcing TVL: the MCMC problem easily converges but show the known and meaningless enlargement of the uncertainty band.
3- Including skewness, enforcing TVL and following the recommendation of Evin et al. (2013): the enlargement disappears (see **Figure 7-Left** in our manuscript, showing the uncertainty band for CRR-GL++ inference)

Evin et al. (2013) partially overcome the methodology followed by Schoups and Vrugt (2010). However, the Generalized Error Model (with the SEP) of Schoups and Vrugt (2010) is more realistic than the Gaussian hypothesis followed by Evin et al. (2013, 2014). Therefore, we have made a sort of amalgamation of both approaches (similar to that tried by Scharnagl et al. (2015)): we infer a full SEP (instead of a skewed Student), plus a variance model, plus an autocorrelation coefficient, all together. And this approach produces the problems shown in Scharnagl et al. (2015), even with the

"easy" FB basin, except when Total Laws are considered. In case of doubts, we have used two known and easily available models, and a basin whose data are free for everyone who wants try it.

With respect the reviewer's comment "…*such a case study catchment cannot provide supporting evidence…*". We are upset that reviewer does not see the results shown in the paper as an evidence of what our research tries to defend: joint inference, as made in recent related papers as the present one, needs force TLs to be statistically correct. We have found several problems with FB basin, as we show in the paper. All of them are consequences of the main problem: TLs are not fulfilled if they are not enforced. Some evidences of this situation are:

A- First set of evidences could be about the effect on the predictive distribution: enlargement and loss of reliability. **Figure 15** shows for GR4J (same hydro model as Evin 2013, 2014) the large difference between GL++ (blue triangle) and GL++NTL. This loss of reliability and resolution when TVL is neglected, is not evidence?
It is true that for CRR model, the difference between GL++ and GL++NTL is not so "explosive" although it exists in resolution (width of the band), as clearly shows the comparison between **Figures 7 and 13**. Both figures have the same Y-axe scale! Why reviewer does not see the enlargement of the uncertainty band, when TVL is not enforced?

B- **Figure 4** shows for both hydrological models CRR (left) and GR4J (right) a different PP-PLOT between GL++ and GL++NTL; in fact, with GR4J is too much different, changing from an overprediction to a strong underprediction (compare solid black line with Total Laws enforcement and the dotted red line without Total Laws). This is a good example, to see how the spurious effects of not enforcing TLs are not always so evident: with CRR the change in the PP-Plot is small, but with GR4J is extraordinary; however, looking at **Figure 12** CRR shows the spurious parameter interaction when TLs are neglected. Therefore, the shown "collateral" effects vary among inferences without TLs, but the root problem is the same.
Perhaps reviewer expected a more "explosive" expansion of the uncertainty band in all cases, but considerations about the size of the enlargement could be subjective in our opinion. We show the results with two models, CRR and GR4J. With GR4J the effects are, in general, more "appealing" than with CRR. However, we wanted include both results because this is a proof that the problem exists (GL++NTL does not fulfill TLs) although the consequences on parameters or predictive uncertainty, strongly depend on the inference problem (even can go unnoticed, as previously shown with the first case study of Schoups and Vrugt (2010)). The low sharpness in which the problem exhibits does not decrease or eliminate the main problem: Total Laws does not fulfill if they are not enforced and this is a statistical pitfall, independently of any other consideration.

C- One more evidence, previously outlined, is in **Figure 12**: Left graphs show for GL++NTL (WLS-AR1 in Evin 2014) the strong interaction between the slope parameter of the error variance model and the autocorrelation coefficient. On the right graph, with the application of TLs, the interaction has disappeared! Interaction is not the cause of the problems; it is another consequence of the real problem. Besides (and the most important), it can be observed the extremely high inferred MAP value for rho (about 0.99) when TLs are neglected. In this case the posterior distribution of rho 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 rho→1, the AR(1) process becomes non-stationary, as explained in Box et al. (1994). However, with the enforcement of TLs, the rho MAP value is about 0.95 (see **Table 2** and **Figure 12**), certainly also very high, but with a perfectly identifiable symmetric distribution and with values lower than 0.96.

Evidences A and B are well explained in the present manuscript, but in the final manuscript we will improve the explanation about Figure 12, with the arguments previously exposed.

Therefore, evidences can be found in the present manuscript, even with an "easy" basin as FB. The joint inference, which includes models for the error variance and for the error dependence, needs enforcing the Total Laws to be correct: this correctness avoids the problematic effects. That is to say: the inference is correct because Total Laws are fulfilled (of course, the other error hypotheses must be also fulfilled). Avoiding the other problems is a "providential" consequence of this. We claim: the way in which we are modeling the errors, through the modeling of its conditional (on qs) distributions, requires their constraining to accomplish with the relation among these conditional distributions of the errors, and the error marginal distribution (see **Figure 1** for a "visual" clarification and the corresponding explanations in **lines 250-270**).

Reviewer's comment #6

As already mentioned above, neither the method nor catchment selection can support such conclusion. But, even if we consider Figure 15 which compares the reliability and resolution of the error models. The WLS error model that the manuscript claims to improve on is clearly amongst the best of the error models under consideration.

Reply

In our opinion, the reliability and resolution (the only good properties of the WLS error model) are not at all the first items in the list of things to be fulfilled by a good error model. For example, the errors dependence (the absence of this) would be more important for considering that an inference is correct. As can be seen in **Figure 5**, WLS model produces an increment in the error autocorrelation, regarding the initial (before the standardization) errors autocorrelation. Precisely we have included WLS (in **Figure 15**) to show how, a bad error model as WLS or SLS, can show good measures of predictive performance, as mentioned in **lines 762-764**.

In this point arise an interesting topic of debate. If the first criteria to judge an inference are the performance metrics about the predictive distribution (as the reviewer seems to claim), and other things as the error hypotheses fulfillment or the plausibility of the inferred hydrological parameters are unimportant, as generally occurs in operational forecasting, then it is not necessary making a joint inference with an appropriate error model. This is explained in **lines 109-132** in the original manuscript. We think (as we explain in those lines) that making inference with an error model deserves to be more exigent with the fulfillment of the error model hypotheses, rather than considering the predictive distribution soundness as the main target. In fact, we account (see for example the **3 last lines** in the abstract) that, being thorough with the error model hypotheses fulfillment leads to get more robust predictive distributions, whereas the opposite never will occur. In short, we propose a course of action which benefits to both the reliability of hydrological parameter estimates and the reliability of the predictive distribution.

We will stress in the final version of the manuscript this point in conclusions.

Reviewer's comment #7

It clearly has other deficiencies, such as lack of treatment of AR1, but this has already been remedied in the literature by including an AR1 term (Evin 2014). So I struggle to see how the conclusions can refer to having addressed problems with this error model. There may be improvements related to the treatment of non-Gaussian errors by including skew and kurtosis in the GL error model, but this has already been shown by Schoups and Vrugt in 2010.

Reply

As we mentioned before, WLS is only included for comparison. Therefore, we discard it as a valid error model only by the fact that it increases the errors dependence (see **Figure 5**). As we also said before, WLS-AR1 of Evin et al. (2014) has its equivalent in GL++NTL of our research, although GL++NTL additionally considers the skewness and kurtosis of the innovations. We agree with reviewer that, our paper improves also to the Evin et al. (2013, 2014) in this sense. We consider through the SEP distribution, the possibility (confirmed also in the paper) that innovations are non-Gaussian.

Of course, Schoups and Vrugt (2010) were the first who included the great idea of using a Generalized Error Model in a Bayesian framework in hydrological modeling, and we have done the proper credits to this in the original manuscript. Nevertheless, as we commented before (Reply #5) they did not include simultaneously all the error model parameters, included in our GL++ inference. They neglected skewness parameter in first case study (FB basin) and the autocorrelation parameter in their second case study (Guadalupe basin). This could seem an insignificant issue. But it is not. Our paper and also the Scharnagl et al. (2015) paper, show how the problems arise in inferences in which all error parameters, including the skewness, are taken into account.

Reviewer's comment #8

4. APPARENT TECHNICAL ERRORS

The discussion on pages 644-657 acknowledges the good performance of WLS to some extent, which helps. However this discussion does not appear in any way anchored to the previous theoretical presentation. To raise these points in the discussion, there has to be a corresponding background presentation of what is a "bad-posed" problem, how is it problematic and how to detect it. The way the manuscript reads at the moment, the study reached unexpected conclusions (which happens) but instead of re-thinking some of its key premises, it tries to "patch" it in the report using concepts that just haven't been properly introduced at that stage of the presentation. Unsurprisingly, erroneous, or least confusing, statements appear to be introduced in this "patch". For example, Line 654 states that in the WLS error model, the same hydro- logical parameter "estimation" is inferred, as well as similar uncertainty bands obtained, using any multiple of (a,k)_ML,TL. Here (a,k) are parameters of the standard deviation of residual errors, sigma=a+k*Q and Q is the streamflow. I really struggle to see how this makes any sense - if we multiply (a,k) by some value c as suggested, sigma will increase by the same factor and the uncertainty bands will inflate accordingly. And the likelihood value will certainly be different. Perhaps this paragraph is missing some major extra detail, or maybe it is a wording issue, or maybe there is some other er- ror/omission in the analysis or calculation, but as it stands it makes no sense. The text states "it can be demonstrated" - please do include the (mathematical) demonstration as it will clarify what you are trying to show.

Reply

We agree with reviewer in which the WLS behavior resulted surprising, even for us. However, we disagree with the consideration of our explanation as a "*patch*". In fact, our explanation is about the "*real patch*".

If we look at **Figure 1-b**, it is logical to have doubts about if TVL is fulfilled or not in WLS without enforce it. Never and nobody (in our knowledge) has applied WLS considering TVL. Generally, WLS has been applied defining the weights of the errors from an error covariance matrix, in a multivariate Gaussian framework. The problem arises if we apply WLS through an error studentization with a model of the conditional (on qs) error variances. In this latter case, it could be possible that inferred conditional variances do not have nothing to do with the variance of the error marginal distribution, differently to what occurs if we work the statistics in a multivariate Gaussian framework in which all probability distributions (conditionals and marginal) are related.

Indeed, we found that WLS (without TLs enforcement and outside of a pure multivariate Gaussian framework) does not fulfill TLs. After discover this issue, we made a WLS inference but enforcing the TVL. Surprisingly, the only thing which resulted appreciably different was the error variance model estimation (for example, see kappa parameter estimation in **Tables 2 and 3**). Even the resulting uncertainty bands were very similar. In short, WLS and WLS+TLs were identical except in two error variance model parameters and in the TLs fulfillment. But, the width of the bands… Why it was not different? Are they theoretically exactly equal? The answer could be named autocompensation. We will try to clarify it, which corresponds to the sentence of our manuscript "*it can be demonstrated*" in **line 652**.

1. In calibration, methods WLS and WLS+TLs, yield similar likelihood and similar hydrological parameters. Why?

We define the residuals or errors as: $e = q - q_s$

We define the error conditional std.dev model as: $\sigma_{e|q_s} = \alpha + \kappa q_s$

We define the variance-stabilized errors as: $\eta = e\sigma_{e|q_s}^{-1}$

We define the std.dev. of the etas as: $\sigma_\eta = \left(V[\eta]\right)^{0.5}$, where $V[\bullet]$ is the variance operator

As our variance-stabilized errors must be ~SEP(0,1,beta,psi) we standardize them:

$a = \eta\sigma_\eta^{-1}$, that is to say, a ~ SEP(0,1,beta,psi)

Now we define a new conditional std.dev model as:

$\sigma'_{e|q_s} = C\sigma_{e|q_s} = C\left(\alpha + \kappa q_s\right) = \alpha' + \kappa' q_s$, with C>0

Obviously, it holds that: $\eta' = \eta C^{-1}$ and also $\sigma'_\eta = \sigma_\eta C^{-1}$

Therefore, the new variance-stabilized standard errors are:

$a' = \eta'\sigma'^{-1}_\eta$, which corresponds to $a' = \eta C^{-1}\sigma_\eta^{-1} C = a$

As a and a' are equal and are the errors which play in the likelihood function, with any of both error conditional std.dev models, the likelihood yields the same results.

2. In simulation of the inferred error model, where the uncertainty bands are generated, occurs a similar compensation effect which yields similar bands in both cases.

We will add this demonstration as one more appendix in the final manuscript.

With respect to the term "*bad-posed*" problem, our manuscript says in **lines 655-656**: "*WLS, in NTL case, is a bad-posed problem to estimate a unique identifiable set of variance model parameters*". In our case, any scaling of the parameter set, which fulfills the TVL, could be valid to get similar inference results. I.e., in WLS inference, as it has been defined, there are infinite "valid" solutions to the heteroscedasticity model inference, although only one of them has statistical sense: the one which fulfills the Total Variance Law.

Reviewer's comment #9

5. THE PRESENTATION IS TOO DISORGANIZED

There is a reason why technical reports have a generally agreed standard set of sections, such as Intro, Theory, Methodology, Results, Discussion, etc. This allows the reader to easily find any details they are interested in. At the moment the specifics of the case study are scattered all across the paper and its very hard to ascertain exactly what experiments were undertaken and why. Please consolidate the presentation and explanation of the methodology in a single section, as this will avoid (likely) confusion on the part of reviewers and readers. Articulating a more precise set of objectives early on, and justifying how the paper will prove these objectives have been achieved, would also help the reader navigate the paper.

Reply

The manuscript in its current form is organized in the following sections:

1- Intro. 2- Generalized error statistical model. 3- Inference. 4- Application to Rainfall Runoff modeling. 5- Results. 6- Discussion. 7- Conclusions

Objectives are stated in **lines 49-60** of the Introduction, and it is not possible to put them earlier. After reviewer's comment we agree that: sections 2 and 3 of the original manuscript could be merged in a new section 2 called Theoretical Framework. Besides section 4 of the original manuscript is the equivalent to a section called Methodology.

For the sake of clarity, we will introduce these two changes in the revised manuscript. At the end of the Introduction section (original manuscript, **lines 73-81**), we tried to explain the structure of the document. We will also improve this explanation in the revised manuscript, adapting it to the new titles. Therefore, this new explanation could be as it follows:

"*The next section, Theoretical Framework, presents the generalized error statistical model and the joint Bayesian parameter inference followed by this research. A new formal generalized likelihood function is presented and the methods for obtaining the posterior of parameters and the predictive distribution are detailed. Section 3, Methodology, describes the different inference settings (one basin, two lumped hydrological daily models and four different error models) used in this paper and how to enforce the Total Laws on them. Section 4, Results, applies the presented theoretical framework to estimate the parameter and predictive uncertainty, through the joint inference of the hydrological and error models described in previous section. The fulfillments of the error model assumptions, as well as several indicators*

of performance are tested in this section. Section 5, Discussion, gives clarifications and points out the relevant results, including the effect of not enforcing the TLs in one error model. Finally, section 6 summarizes our findings and conclusions."

Regarding the affirmation "*specifics of the case study are scattered all across the paper*". We disagree. The case study (basin, hydrological models and inferences) are only presented in the present section 4 (section 3 in the revised manuscript).

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C.: Time Series Analysis: Forecasting & Control., 1994.

Evin, G., Kavetski, D., Thyer, M. and Kuczera, G.: Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration, Water Resour. Res., 49(7), 4518–4524, doi:10.1002/wrcr.20284, 2013.

Evin, G., Thyer, M., Kavetski, D., McInerney, D. and Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, Water Resour. Res., 50(3), 2350–2375, doi:10.1002/2013WR014185, 2014.

Scharnagl, B., Iden, S. C., Durner, W., Vereecken, H. and Herbst, M.: Inverse modelling of in situ soil water dynamics: accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals, Hydrol. Earth Syst. Sci. Discuss., 12(2), 2155–2199, doi:10.5194/hessd-12-2155-2015, 2015.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res., 46(10), 1–17, doi:10.1029/2009WR008933, 2010.