First of all, we would want to thank the two anonymous referees and Dr. Tyralis for their suggestions, comments and questions. We are sure they have been useful to improve the manuscript.

We are aware that changes are required in the manuscript, as we pointed out in our replies. However, referees highlighted the good research quality of the manuscript and they found it relevant. Significant number of downloads of the manuscript, from HESS web page, is also a good indicator that our research generates interest in the community. It is always a pleasure to hear that your work is receiving attention and interest and is providing new insights. Now, we are going to summarize the main raised research comments and our replies.

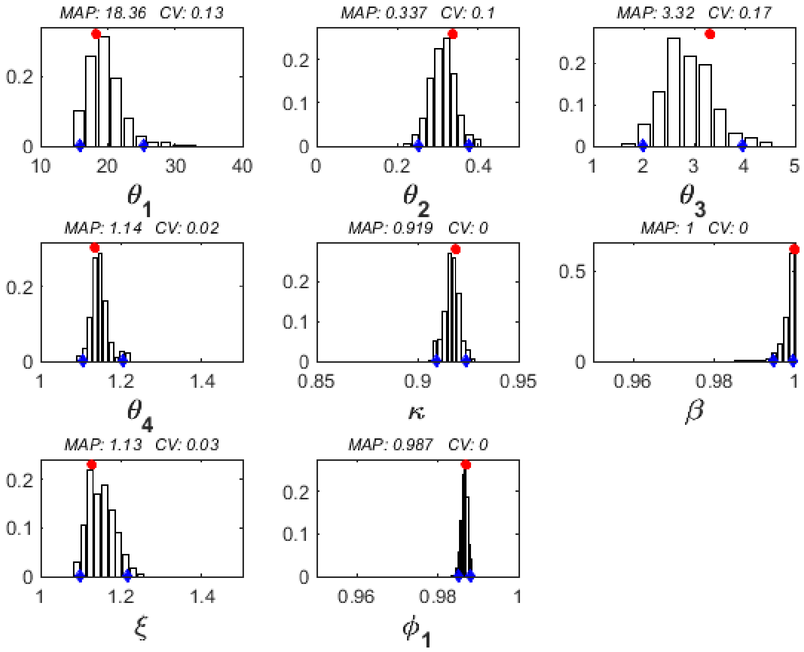
Firstly, we want to clarify the main target of the paper. Our research has always been within a Bayesian calibration framework and not on “operational hydrology”. This means that, obtaining the best possible reliable parameter set is mandatory for us, hence we only consider the **joint inference** option, namely, inferring all error model parameter jointly with all hydrological ones. Achieving this parameter reliability is possible through an error model which acceptably fulfills all its hypotheses. Moreover, with the compliance of all errors’ hypotheses, namely having a reliable error model, we are also able of reliably assessing the hydrological model predictive uncertainty, since the predictive uncertainty bands are built from the error model.

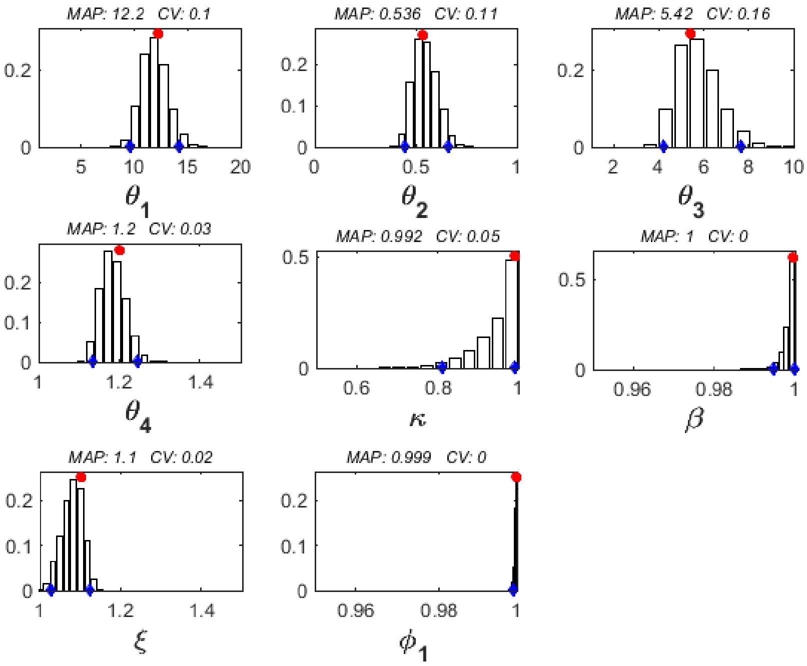
Therefore, the aim of the research was to emulate previous related papers, mainly research of Schoups and Vrugt (2010) with the modifications proposed by Evin et al. (2013). However, from the beginning of our research, we came across **unexpected problems**. The fact is that, when in a joint inference, the error changing-variance (and/or the changing-bias) is modeled as in Schoups and Vrugt (2010), Evin et al. (2013, 2014), Scharnagl et al. (2015) or ourselves, Total Laws (TLs) enforcement is necessary for their compliance. And compliancy of TLs should not be optional from a strict statistical point of view. From the onset of those unexpected problems, we did not want to find the best error model (neither the best hydrological model) which yields the best performance for the French Broad basin. Hypothesized components of the error model (variance model, bias, model, etc.) were chosen with the main purpose of conducting the true reason behind our manuscript: bringing to the light the statistical requirement of the **Total Laws’ fulfillment** and making the recommendation of having particular care about it, when joint inferences with sophisticated error models are performed. Of course, fulfillment of TLs is not enough to achieve reliable inferences, but it is a statistically necessary condition. Besides, suitability of hydrological and error models is also essential in the achievement of this reliability.

Secondly, in response to the question requested by first reviewer, about the necessity of using a different more “problematic” basin, our opinion is that we are presenting a mathematical (statistical) problem, and we are dealing with it and proposing one solution, from a mathematical point of view. We have also demonstrated empirically the suitability of our solution with the French Broad basin. There is **one single mathematical result** which evidences that something goes wrong by neglecting TLs in inferences with sophisticated error models, as for example with WLS or GL++:

after its inference, calculation of error (marginal) variance yields a result which does have nothing to do with which would be mathematically expected, that is to say, with the mean of the conditional variances. But, in addition to this mathematical issue, **more examples of another different spurious problems** were shown in the manuscript and/or posted in replies to reviewers, concerning different issues: non-identifiability of error model parameters, meaningless enlargement of the uncertainty band or even the abnormal increment of the correlation between hydrological parameters, shown all of them when TLs are neglected with complex error models (e.g. GL++, GL++bias).

Even so, during this discussion, we have made an important additional effort trying to contribute with another clarifying example of inference, with another different basin. In this case, we have chosen a “problematic” basin which is also within the MOPEX project, with name Guadalupe River. It is the driest basin in the MOPEX experiment, and its challenge was demonstrated in Evin et al. (2014) among many others. Serving as example of how difficult can be its modeling, the GR4J’s calibration with SLS error model yields a low Nash-Sutcliffe index (NSE=0.46), which indicates the existence of severe problems in the hydrological model structure and/or in the data. We have performed **the Bayesian joint inference of GR4J’s parameters, with the GL++ error model, for Guadalupe basin**, and for the same period considered with French Broad basin. Unsurprisingly, results do not add new important or significant things, that we had not previously shown for the French Broad basin. As an example, in the following figure we show the inferred posteriors with TLs enforcement (GL++ on top) and without it (GL++NTL on bottom).





There are two main things to remark about these two new figures, and both are in the same sense. Firstly, ***kappa* parameter**, the slope in the linear variance model is identifiable (although it has a high value) when TLs are enforced, but it is not when TLs are neglected (adopts the value of one, which is the established higher bound for this parameter). It is important to note that **this is a new spurious effect**, which did not appear in any previous inference with French Broad basin. Secondly, **autocorrelation parameter *phi*$_1$** is identifiable (although also with a very high value) when TLs are enforced, but it is not when TLs are neglected, because it adopts the value of one, which does not make sense in an AR(1) model. Effectively, extremely high values for *kappa* and *phi*$_1$ are a symptom that Guadalupe is a “difficult” basin (surely as many others) and the generated error structure seems too complex, to be modeled with a simple linear variance model and a first-order autoregressive model, as we and previous researches have made. However, we have been able to identify the full parameter set, but only when we have enforced TLs.

Thirdly, in response to the question, why do not problems appear in all MOPEX basins? Schoups and Vrugt (2010) introduced an innovative and generalized formal Bayesian aggregated approach for the error modeling. This sophisticated, but also highly flexible approach is based on the independent modeling of practically all the statistical features of the error conditional distributions: its variance, bias, kurtosis and skewness. Therefore, the Schoups and Vrugt (2010) methodology allows to **characterize the heteroscedasticity, the bias and non-normality** of the hydrological model errors.

So far, the **treatment of the error heteroscedasticity** had been performed through a previous transformation (e.g. **Box-Cox transformations**) of the original variables, with the aim of stabilizing the non-constant errors variance. Research of Del Giudice et al. (2013) is a good and recent example among many others. Long time ago, Sorooshian and Dracup (1980) were the first who made use of a **direct method** (considering the untransformed original variables) for the error changing-variance modeling. This is a special case of a direct error variance modeling, because their direct method was partially based on a constant error variance, previously obtained from a stabilizing transformation. On balance, their direct methodology established "special" conditions (or restrictions) on the error variance model parameters, for their inference with the untransformed original variables. That is to say, **error variance model parameters were jointly but not freely inferred** with the hydrological ones. After that work and during a prolonged period, nobody (in our knowledge) tried a direct approach for the inference of an error variance model.

Yet recently, the proposed approach by Schoups and Vrugt (2010) also adopted the direct treatment for the error variance modeling. But in their methodology, with a simple linear error variance model (same as ours and as Sorooshian and Dracup (1980)), they did not impose any condition to infer these variance model parameters, differently to Sorooshian and Dracup (1980). Schoups and Vrugt (2010)'s approach, was based on the model specification for the statistical features of the error conditional distributions. In this respect, one of our first questions was, how these error conditional distributions, with practically all its properties varying during the inference, can build their parent bivariate pdf P(e-q_s) distribution? Our belief is that some kind of additional formulation (in essence, a sort of **restriction in the error model parameters**) could be necessary for maintaining the statistical coherence of the whole. When this statistical coherence occurs, Total Laws are fulfilled, or what is the same, **marginal and conditional distributions** are properly related, since both **belong to a unique properly defined bivariate distribution**.

Our opinion is that, the greater or lesser **complexity of the inferred error structure** (mathematically represented by the bivariate pdf P(e-q_s)) seems to predetermine the **occurrence (or not) of the spurious problems** during the joint inference of the variance model, when its parameters are not properly constrained. French Broad basin as an "easy" basin, it has an error structure not too much complex and the problems only appear when 5 unrestricted error model parameters are freely inferred, as shown in our original manuscript: however, as it was shown by Schoups and Vrugt (2010) and in our manuscript (with the restriction of one of the variance model parameters by fulfilling TLs), with 4 freely inferred parameters (and of course also with 3, in Evin et al. (2014)), this basin does not show problems, even if TLs are not considered. On the other hand, basins which are more difficult to model, as Guadalupe basin, yield a more complex error structure whose joint inference does not support, without showing problems, to leave free the variance model parameters. Guadalupe basin showed problems in Evin et al. (2014), with only 3 freely inferred error model parameters (2 for the variance and 1 for the AR (1)). In our inference on Guadalupe, we have not found spurious problems with "five" error model parameters (truly four, if we consider the restriction imposed by TLs for one of the variance model parameters); but as could be expected, since Guadalupe had problems yet with only 3 freely inferred parameters, with five free parameters, we have also found problems. Therefore, the question could be, how many error parameters supports the inference of a basin, without showing problems, when its error variance model parameters are freely inferred, namely without TLs enforcement? In our opinion, this question does not make sense from a statistical point of view, and it is out of the scope of our paper.

From Schoups and Vrugt (2010)'s, other researches who have performed their same approach for the modeling of the error variance (e.g. Evin et al. (2013, 2014), Scharnagl et al. (2015)), have neglected the statistical necessity of relating marginal and conditional error distributions. This is a statistical pitfall: we realize that this is a hard affirmation, which inevitably does not sound good. But sincerely, we do not find how to word this in a better but also unambiguous manner. Another different thing is that, by neglecting the TLs issue, problems appear or not in the inference process. The underlying statistical incoherence exists by only not fulfilling TLs, although spurious problems do not appear. We recognize the **non-existence of a biunivocal relation** between violation of TLs and the occurrence of problems with the inference, as previous researches have demonstrated, and we have also tested. However, the inexistence of this reciprocity does not mean that those inferences without problems are correct. We do not say that, necessarily they are not correct. We only say that **to be statistically correct, they should comply with TLs**, if not by its own, through some other restrictions or procedure. Our proposed procedure is the TLs enforcement, as explained in our manuscript.

Total Laws could be understood as one method to reduce the degrees of freedom in the inference problem. However, in our opinion, TLs enforcement is more than an ad-hoc method to restrict the error model parameters. Meeting TLs is, from the theoretical point of view, a statistical requirement which eventually produces the convenient error model parameter restriction. Total Laws are fulfilled in classical inferences (as the SLS method) and we understand that, compliance of these laws should be transposed to any inference which involves to the inferred variables and its errors.