Responses to all the Referees:

We sincerely appreciate the comments and advices from the Editor and Referees. Detailed responses to the comments from the Editor and Referees are presented below. In the following Responses, comments by Referee #1 are labeled A, and comments by Referee #2 are labeled B. For example, A1 represents the first comment made by Referee #1 and B1 represents the first comment made by Referee #2. Please also note that the page and line numbers mentioned in reviewers comments refer to the original version, while in the authors' response they refer to the revised version.

Responses to Editor:

The paper is potentially valuable, however, as the reviewers pointed out, there are several points of concern. One of the aspects that did not convince me is the combination of heuristic and Bayesian objective functions in Equation 5.

I suggest the author to take into serious consideration the various comments before resubmitting their paper.

Reply: We sincerely appreciate the comments from the Editor. Great efforts have been made to address the editor and reviewers' comments. The equation.5 has been modified in the revised manuscript, and more information has been provided in the revised manuscript. Please refer to lines 281-291, pages 14-15.

Responses to Referee #1:

This paper analyzes the prediction performance of a lumped hydrological model using different time and spatial dependent parametrizations of one of its parameters. There are several errors in the paper and points that should be explained better and I have a major concern regarding the results.

Comment on the results:

A1: The value of omega looks strange to me. Assuming that the equation 1 you wrote is correct (and therefore it is a frequency and not a phase) and that the order of magnitude of omega is of hundreds (like shown in figures 8 and 9), this mean that your parameter theta1 oscillates hundreds of times per time step. This looks unreal to me since the goal of having time-variant parameters is to represent long term (seasonal) oscillations. Therefore, either there is a problem with the unit of omega or your model is not doing what it was meant for. If omega is a phase (meaning theta1 = alpha + beta*sin(t + omega)) the value of omega makes more sense but theta1 would still complete an oscillations every 6.28 time steps (the time step is days, right?). Don't you also have a frequency that multiplies "t" and have a small value?

Reply: We deeply apologize for our mistakes.

(1) ω represents frequency rather than phase. It has been revised accordingly in the revised manuscript. Please refer to line 222 in page 11 and line 520 in page 26.

(2) We have carefully checked the data for Figures 8 and 9, and found that we misused the results of regression parameter α to plot for ω. We are so sorry for our mistakes. As is stated in the response to comment A18, the Figures 8 and 9

have been modified as violin plots with the right data. Figures are as follows:

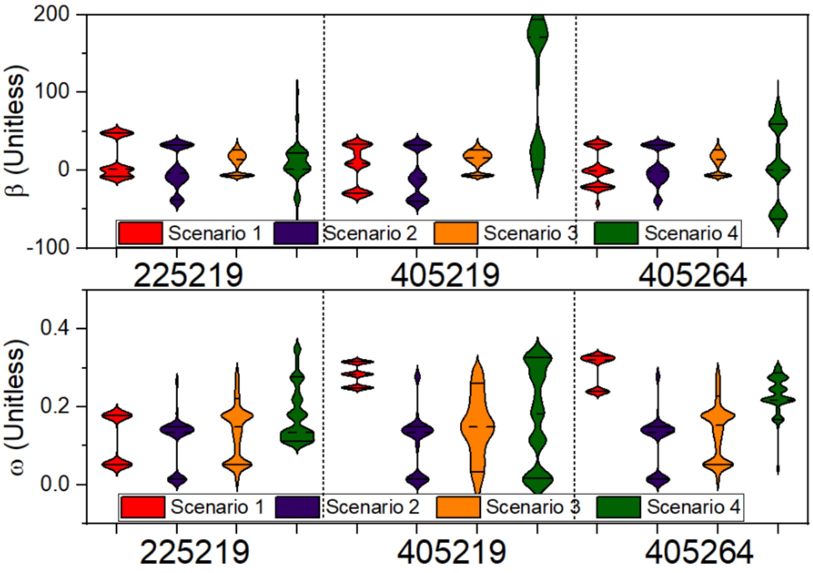

Figure 8. Posterior distributions of the regression parameters (β and ω) for the production storage capacity (θ_1) for the four modeling scenarios in all the 3 studied catchments. In this figure, parameters were calibrated in the non-dry period while verified in the dry period. The solid horizontal lines within the violin plots denote the 25th and 75th percentiles of the posterior distribution, while the dash line denotes median estimates.

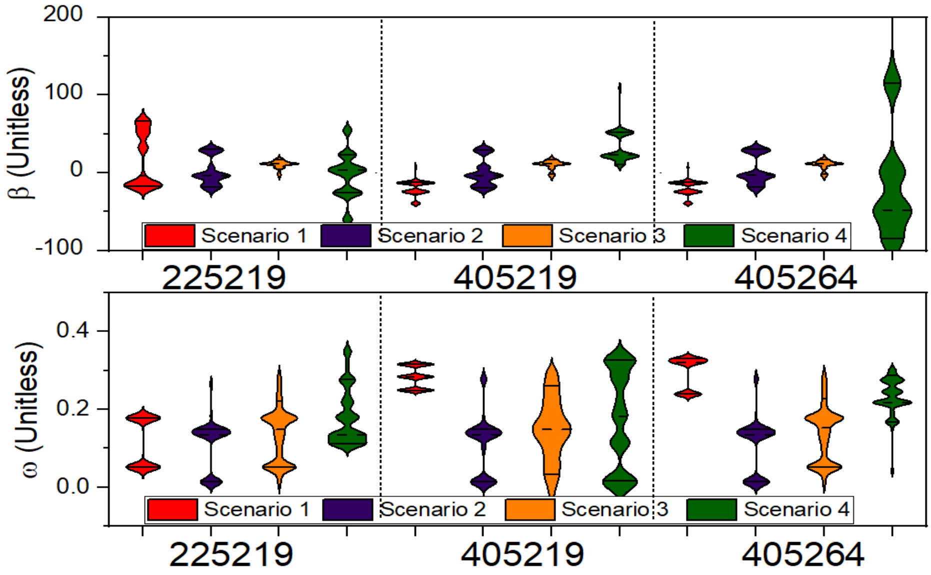

Figure 9. Posterior distributions of the regression parameters (β and ω) for the production storage capacity (θ_1) for the four model scenarios in all 3 studied catchments. In this figure, parameters were calibrated in the dry period while verified in the non-dry period. The solid horizontal lines within the violin plots denote the 25th and 75th percentiles of the posterior distribution, while the dash line denotes median estimates.

As shown in Figure 8, the catchment averages of regression parameter ω for different scenarios are 0.24, 0.14, 0.15, and 0.18, while those in Figure 9 are 0.15, 0.26, 0.23, and 0.17 respectively. The phase T of the sine term could be derived based on the estimates of ω based on equation $T = 2\pi/\omega$. Thus, the mean phases T of model parameter θ_1 for different scenarios are 26.2, 46.3, 41.9 and 35.2 in Figure 8, respectively. Similarly, the mean phases T are 42.9, 24.1, 27.4 and 38.0 in Figure 9, respectively. Please refer to lines 521-527 in pages 26-27.

Detailed comments:

A2: line 102-103: There is not a clear definition of pooling, complete pooling and hierarchical Bayesian. I would explain shortly what do they mean and which are the differences since then the paper only writes about hierarchical Bayesian.

Reply: Thank you for your comments. The following explanations about the no pooling, complete pooling and hierarchical Bayesian framework have been added in the revised manuscript:

In general, there are three methods to consider the spatial coherence between different catchments in parameter estimation. The first one is no pooling, which means every catchment is modeled independently, and all parameters are catchment-specific. The second one is complete pooling, which means all parameters are considered to be common across all catchments. The third/last one is hierarchical Bayesian (HB) framework, also known as partial pooling, which means some parameters are allowed to vary by catchments and some parameters are assumed to be drawn from a common hyper-distribution across the region that consists of different catchments. Please refer to lines 99-107 in page 5.

A3: line 152-153: It would be beneficial to explain shortly how the method works even if it was already used in other studies.

Reply: Thank you for your comment. Definition of the dry period is explained as follows and has been added in the revised manuscript:

Saft et al. (2015) tested several algorithms for dry period delineation, which considered different combinations of dry run length, dry run anomaly and various boundary criteria, and found that the identification results of dry period by one of the algorithms showed marginal dependence on the algorithm and the main results were robust to different algorithms. The detailed processes could be found on Saft et al. (2015) and are also presented as follows.

Firstly, the annual rainfall data were calculated relative to the annual mean, and the anomaly series was divided by the mean annual rainfall and smoothed with a 3 year moving window. Secondly, the first year of the drought remained the start of the first 3 year negative anomaly period. Thirdly, the exact end date of the dry period was determined through analysis of the unsmoothed anomaly data from the last negative 3 year anomaly. The end year was identified as the last year of this 3 year period unless: (i) there was a year with a positive anomaly >15% of the mean, in which case the end year is set to the year prior to that year; or (ii) if the last two years have slightly positive anomalies (but each <15% of the mean), in which case the end year is set to the first year of positive anomaly; (iii) to ensure that the dry periods are sufficiently long and severe, in the subsequent analysis, the authors use dry periods with the following characteristics: length ≥ 7 years; mean dry period anomaly<25%.

Please refer to lines 159-176 in pages 8-9 in the revised manuscript.

A4: line 159: Maybe it is more appropriate to use "cross validation" instead. I suggest to avoid making a paragraph with just one sentence and remove paragraphs 2.1.1 and 2.1.2 putting all together in section 2.1.

Reply: Thanks. Following the Referee's suggestion, paragraphs 2.1.1 and 2.1.2 have been put together within section 2.1, and the sub-titles of sections 2.1.1 and 2.1.2 have been deleted in the revised manuscript. Please refer to lines 150-183 in pages 8-9 in the revised manuscript.

A5: chapter 2.3: It is not clear to me what do you do with the other parameters of the GR4J model (theta2, theta3, theta4). Do you keep them fixed or do you sample them? What is their effect on the final result?

Reply: Thank you for your comments.

(1) All other model parameters (θ_2, θ_3, and θ_4, except θ_1) are not fixed, but sampled simultaneously with regression parameters α, β and ω (if present), and hyper-parameters μ_2, σ_2, μ_3 and σ_3 in the SCEM-UA algorithm. In real calculation process, we would set a large variation interval for each unknown quantity first, the estimations of parameters would converge to a smaller interval in MCMC calculation process. Then we checked the model convergence using the Gelman-Rubin convergence value by evolving three parallel chains with 30000 random samples and confirmed that the convergence value was smaller than the threshold 1.2 (Gelman et al., 2013).

Please refer to lines 285-291 and lines 295-298 in page 15 in the revised manuscript.

(2) Previous studies on GR4J model showed that θ_2, θ_3, and θ_4 are less sensitive than θ_1 under changing climate (Perrin et al., 2003; Renard et al., 2011; Westra et al., 2014). Therefore, we think that it is reasonable to assume that θ_1 is time-varying while other model parameters are temporally invariant. Please refer to lines 206-217, pages 10-11.

A6: line 199: The equation is different from the ones reported in Table 1.

Reply: We apologize for our mistakes. The fault equation in Table 1 has been revised as equation 1 in the revised manuscript. Please refer to Table 1 in the revised manuscript.

A7: line 201: You write that omega is the phase while in the equation 1 it is a frequency.

Reply: Thank you for pointing out this mistake. The ω represents the frequency rather than the phase (see response to comment A1). The statement in line 201 in the original manuscript (see line 222, page 11 in the revised manuscript) is wrong and has been modified in the revised manuscript. Please refer to line 222, page 11.

A8: line 202: The combination alpha=beta=omega=0 makes theta 1 to be equal to 0, that indeed it is a constant value but probably it is not what you want.

Reply: Thanks. According to the definition of the GR4J model (Perrin et al., 2003), θ_1 represents the primary storage of water in the catchment and must be a positive value. Thus, in order to satisfy this requirement, the combination of $\alpha=\beta=\omega=0$ would be excluded under calculation, and other combinations that made θ_1 equal to zero would be excluded too.

A9: chapter 2.3.2: What happens to alpha? You don't write about it anymore in the rest of the paper. Do you keep it fixed or do you sample also it? What is its effect on the final result?

Reply: Thanks.

(1) The regression parameter α represents the constant term in equation 1. Changes in α lead to consistent changes in θ_1 across the whole time series, which doesn't result in temporal variations of model parameter.

(2) Regression parameter α is not fixed in advance but is sampled as same as other unknown quantities. The posterior distribution of α is derived out simultaneously with hyper-parameters μ_2 , μ_3 , σ_2 and σ_3 , other regression parameters β and ω (if present), and model parameters θ_2 , θ_3 and θ_4 in the SCEM-UA algorithm. Please refer to lines 295-298, page 15.

A10: chapter 2.3.2: It is not clear to me if linking the parameters between catchments means sampling them from the same Gaussian distribution or there is another form of linking.

Reply: We are sorry for not making this part clear enough. The link is that regression parameter β (or ω) of different catchments is assumed to sample their values in the same Gaussian distribution. This kind of links has been widely used in the field of extreme event analysis, such as Sun et al. (2015, 2016), Lima et al. (2009) and Bracken et al. (2018).

A11: chapter 2.3.2: How do you sample omega and beta when they are not linked?

Reply: Thanks. ω is not linked in scenario 1, while β is not linked in scenario 2. In scenario 4, both ω and β are not linked. Spatially irrelevant parameters of different catchments would be sampled and derived as independent variables. For example, In scenario 1, regression parameter β is spatially linked, i.e., $\beta(c) = N(\mu_2, \sigma_2^2)$, which means that the estimates of β are shared by all catchments. Meanwhile, independent regression parameters $\omega_{1\text{-}1}$, $\omega_{1\text{-}2}$, and $\omega_{1\text{-}3}$ are used as independent variables to represent the frequency of model parameter θ_1 in different catchments. The name of all unknown quantities and their prior ranges in different scenarios could be found in the supplementary material.

Please refer to lines 254-264, page 13.

A12: line 218: How do you choose the values of mu and sigma, the hyper-parameters of your model?

Reply: Thanks. The posterior distributions of all unknown quantities, including model parameters θ_2, θ_3 and θ_4, and regression parameters α, β and ω, and hyper-parameters μ_2 , μ_3, σ_2 and σ_3 are sampled and derived simultaneously through the SCEM-UA algorithm. In actual calculation process, we would set a large variation interval for each unknown quantity first, parameters would converge to a small interval in MCMC calculation process. Then we checked the model convergence using the Gelman-Rubin convergence value by evolving three parallel chains with 30000 random samples and confirmed that the convergence value was smaller than the threshold 1.2 (Gelman et al., 2013). Please refer to lines 295-298 in page 16 and lines 306-312 in page 16.

A13: chapter 2.4.1: I wouldn't call "likelihood function" what actually is an objective function.

Reply: Thanks. As suggested, the "likelihood function" has been modified as "objective function" in the revised manuscript. Please refer to lines 268, page 14.

A14: line 250: You are mixing an objective function with a prior distribution of the parameters. How do you account for the prior distribution of the parameters when they are not linked?

Reply: Thanks. The uniform distribution is used as the prior distribution for all unknown quantities.

(1) The objective function of Eq.1 has been modified as follows:

$$\varepsilon_c\left[\theta_1,\theta_2,\theta_3,\theta_4\right] = \text{-}RMSE\left[\sqrt{Q}\right]\left(1+\left|1+BIAS\right|\right) \tag{1}$$

where

$$RMSE\left[\sqrt{Q}\right]=\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[Q_{sim}\left(t\right)-Q_{obs}\left(t\right)\right]^{2}} \tag{2}$$

and $RMSE\left[\sqrt{Q}\right]$ refers to the root-mean-square error, in which Q_{sim} is derived by the adopted hydrological model.

(2) The objective function of Eq.5 has been modified as follows:

$$\begin{aligned}
&Scenario\ 1:\ \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta,\omega(c)\right]\bullet f_N\left(\beta\middle|\mu_2,\sigma_2\right)\\
&Scenario\ 2:\ \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta(c),\omega\right]\bullet f_N\left(\omega\middle|\mu_3,\sigma_3\right)\\
&Scenario\ 3:\ \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta,\omega\right]\bullet\prod_{n=1}^{2}f_N\left(\beta,\omega\middle|\mu_2,\sigma_2,\mu_3,\sigma_3\right)\\
&Scenario\ 4:\ \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\right]\\
&Scenario\ 5:\ \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(c),\theta_2(c),\theta_3(c),\theta_4(c)\right]
\end{aligned} \tag{5}$$

where the number of catchments in the region is represented by C; *c* represents the specific catchment; the *t* is the time step.
Please refer to lines 281-291, pages 14-15.

A15: chapter 2.4.2: You don’t say which settings of the sampling method you use (e.g. how many parameters you sample. . .)

Reply: Thanks. The sampling method used in this paper is the SCEM-UA algorithm. The detailed descriptions of the settings of SCEM-UA algorithm have been added in the revised manuscript:

(1) Convergence is assessed by evolving three parallel chains with 30000 random samples, the posterior distributions of parameters are evaluated by the Gelman-Rubin convergence value and are confirmed that the convergence value is smaller than the threshold 1.2 (Gelman et al., 2013). Please refer to lines 306-312, page 16.

(2) The number of unknown quantities in different scenarios is as follows: fifteen in scenarios 1 and 2, thirteen in scenario 3 and eighteen in scenario 4. Please refer to lines 261-262, page 13.

A16: chapter 3.2.1: The dataset that you get is unbalanced, since there are more wet years. Is it taken into account? Does it have an effect on the calibration?

Reply: Thank you for pointing out this situation. Generally, a longer time series may improve the robustness of hydrological predictions. However, we tested the calibration performance with different lengths of records (> 6 years) in dry (non-dry) period and found that their results are almost the same. Therefore, we used the whole time series of the dry (15 years) and non-dry (10 year) periods into model calibration.

A17: chapter 3.2.3: Figures 7 and 8 are actually 8 and 9.
Reply: Thanks. Changes have been made.

A18: Figures 5, 6, 8, 9: Since you want to show a probability distribution I wouldn’t use a boxplot but, instead, I suggest to use a violin plot (e.g.https://seaborn.pydata.org/examples/grouped_violinplots.html)
Reply: Thank you for your suggestions.

(1) As suggested, Figures 8 and 9 have been modified as violin plot in the revised manuscript, which also could be found in response to comment A1 by Referee #1

(2) Figures 5 and 6 have been revised as violin plot in the revised manuscript.

A19: Figures 8, 9: Why do you change the colors between beta and omega? This makes the plot more difficult to read.
Reply: Thanks. Changes have been made as suggested.

Responses to Referee #2:

General Comments

The study of Pan et al. tests a Hierarchical Bayesian framework to incorporate time and spatial variability in model parameters. Specifically, the method was tested for the GR4J-model in three Australian catchments. Four modelling scenarios were tested, and one base scenario was formulated. The study shows that including spatially and temporally variable parameters improves model performance and reduces uncertainty. The article shows interesting work, which could be a nice contribution to the field.

Generally, the article needs some more explanations on the method, but there is also some incomplete reasoning. Hence, there are several issues I'd like to address.

Specific comments

B1: Key of the article is the hierarchical framework, but the authors may want to work on the explanation of the method. It is especially not clear to me how the hyper-parameters are determined, and how the catchment-specific values follow from that. Are the hyperparameters estimated in SCEM-UA? Or are these pre-defined? The gaussian distributions are defined by the authors as prior distributions, and that makes me assume that the model parameter theta is determined in SCEM-UA starting from this prior distribution, whereas the remaining model parameters are either kept fixed or sampled from a uniform distribution and independently for each catchment. Is that correct? Because if that is the case, the hyper-parameters (and hence the distribution) are determined in advance, so what are these based on? Besides, the choice of a gaussian distribution may seem a logical first guess, but it remains an arbitrary choice. So what is the reasoning behind this choice? In addition, the choice of the prior distribution may lead to some circular reasoning. When spatial coherence is used, the variation in performance goes down, but is this not just an artefact of the pre-defined gaussian distribution? In other words, if the prior distribution is set narrower, the resulting posterior distribution will probably be narrower as well. I believe it is therefore crucial to report also the prior ranges (or fixed values) for especially the (time-invariant) theta-parameter, but also all other model parameters.

Reply: Thank you for your comment. Since that several sub-comments have been included in comment B1, for clarification, a point by point response to these sub-comments is made as follows. For example, B1S1 refers to the first sub-comment in B1.

B1S1: Key of the article is the hierarchical framework, but the authors may want to work on the explanation of the method. It is especially not clear to me how the hyper-parameters are determined, and how the catchment-specific values follow from that. Are the hyperparameters estimated in SCEM-UA? Or are these pre-defined?

Reply: Thanks for the comments and sorry that we failed to describe it clear enough in the original submission. It is explained here and similar clarification has been made in lines 295-298 in page 15 and lines 306-312 in page 16 in the revised manuscript, and supplementary material.

(1) All the hyper-parameters are not determined in advance. The

hyper-parameters are sampled and determined simultaneously with other unknown quantities in the SCEM-UA algorithm. Actually, all other model parameters (θ_2, θ_3, and θ_4, except θ_1) are sampled simultaneously with regression parameters (α, β and ω (if present)) and hyper-parameters (μ_2, σ_2, μ_3 and σ_3) in the algorithm. Convergence is assessed by evolving three parallel chains with 30000 random samples, the posterior distributions of parameters are evaluated by the Gelman-Rubin convergence value and are confirmed that the convergence value is smaller than the threshold 1.2 (Gelman et al., 2013).

(2) Regression parameter ω is not linked in scenario 1, while β is not linked in scenario 2. In scenario 4, both ω and β are not linked. It should be noted that the spatially irrelevant parameters would be sampled and derived as independent variables. For example, in scenario 1, regression parameter $\beta(c) = N(\mu_3, \sigma^2)$, which means that β is shared by linked catchments, while regression parameters $\omega_{1\text{-}1}$, $\omega_{1\text{-}2}$, and $\omega_{1\text{-}3}$ are used as independent variables to represent the frequency of model parameter θ_1 in different catchments. The names of all unknown quantities in different scenarios have been added in the supplementary material.

B1S2: The gaussian distributions are defined by the authors as prior distributions, and that makes me assume that the model parameter theta is determined in SCEM-UA starting from this prior distribution, whereas the remaining model parameters are either kept fixed or sampled from a uniform distribution and independently for each catchment. Is that correct? Because if that is the case, the hyper-parameters (and hence the distribution) are determined in advance, so what are these based on? Besides, the choice of a gaussian distribution may seem a logical first guess, but it remains an arbitrary choice. So what is the reasoning behind this choice? In addition, the choice of the prior distribution may lead to some circular reasoning. When spatial coherence is used, the variation in performance goes down, but is this not just an artefact of the pre-defined gaussian distribution? In other words, if the prior distribution is set narrower, the resulting posterior distribution will probably be narrower as well. I believe it is therefore crucial to report also the prior ranges (or fixed values) for especially the (time-invariant) theta-parameter, but also all other model parameters.

Reply: Thank you. It is explained here and similar clarification has been made in the revised manuscript.

(1) The Gaussian distribution is one of the widely used distributions for describing the prior layer within the HB framework and has been applied in many previous studies, such as Sun et al (2015, 2016) and Chen et al (2014). The choice of the distribution is not our key point, so we just adopt a typical one from historical

literature. Please refer to lines 287-291 in page 15.

(2) Only two things were fixed in advance: the structure of Gaussian distribution and the variation ranges of all unknown quantities. The posterior distributions of all unknown quantities would be derived simultaneously in the SCEM-UA algorithm. In addition, as illustrated in response to comment A14 by Referee #1, convergence is assessed by evolving three parallel chains with 30000 random samples. The posterior distributions of parameters are evaluated by the Gelman-Rubin convergence value and are confirmed that the convergence value is smaller than the threshold 1.2 (Gelman et al., 2013). Please refer to lines 306-312 in page 16.

(3) The prior ranges of all unknown quantities in different scenarios have been added in the supplementary material.

B2: I also wonder how valid it is to assume the catchments are similar. The authors state on p15.L314, that the catchments satisfy the homogeneity assumption. What is this assumption and how do they satisfy this assumption? A clear description of the catchments may be needed to defend that the catchments are the same. Just looking at the DEM and the annual values of rainfall and runoff (Table 2) give me the idea that the Big catchment (405264) behaves fundamentally different compared to the other two. This catchment also reached much higher performances in calibration (Fig.5 and 6) when no spatial coherence is used, and also shows different results in the BIAS comparison (Figure 7).

Sometimes, the conclusions and statements of the authors do not seem to be strongly supported by the data as shown. The boxplots with performances (Figures 5, 6) show relatively similar performances, and, to be honest, a clear pattern is not very obvious. In addition, the authors tend to generalize in some cases findings that mainly apply to just two of the three catchments (see also my minor comments). I believe additional analyses may be needed to support the conclusions more, for example a statistical test to check if the distributions are significantly different. Or the addition of other, multiple performance measures, to assess the performance over multiple aspects (high flows, low flows etc.). Further, all beta-values plot around zero in Figure 8, basically pointing at the absence of a clear trend. Is this indeed true? It would be interesting to show the timeseries of the parameter. The absence of a trend may explain the similar performances for all scenarios, and especially also why the time-varying scenarios do not outperform the others clearly. Besides, when beta is around zero, there is no point of looking at omega, as this does not do much in that case.

Concluding, the authors may need to clarify more what they did and how they arrive at several conclusions. I hope the authors find my comments useful, and I look forward to a revised manuscript.

Reply: Thank you for your insightful comments. Since that several sub-comments have been included in comment B2, for clarification, a point by point response to these sub-comments is made as follows.

B2S1: I also wonder how valid it is to assume the catchments are similar. The authors state on p15.L314, that the catchments satisfy the homogeneity assumption. What is this assumption and how do they satisfy this assumption? A clear description of the catchments may be needed to defend that the catchments are the same. Just looking at the DEM and the annual values of rainfall and runoff (Table 2) give me the idea that the Big catchment (405264) behaves fundamentally different compared to the other two. This catchment also reached much higher performances in calibration (Fig.5 and 6) when no spatial coherence is used, and also shows different results in the BIAS comparison (Figure 7).

Reply: We apologize for our mistakes. The homogeneity assumption has be deleted in the revised manuscript, because it is the spatial coherence of adjacent catchments that has been used as effective information to restrict the prediction uncertainties, so the homogeneity assumption of different catchments is not necessary. The studied catchments do have several similar characteristics: i) the average slope is similar, that is, catchment 225219 is 12.8, catchment 405219 is 10.7 and catchment 405264 is 9.7; ii) as shown in Table 2, these catchments have similar climatic conditions including mean annual potential evapotranspiration and rainfall patterns; iii) these catchments have experienced the same prolonged drought, and have similar amplitude of variation of rainfall and runoff between non-dry and dry periods.

B2S2: Sometimes, the conclusions and statements of the authors do not seem to be strongly supported by the data as shown. The boxplots with performances (Figures 5, 6) show relatively similar performances, and, to be honest, a clear pattern is not very obvious.

Reply: We agree with the Referee that Figures 5 and 6 showed similar pattern in terms of the ranked orders of NSE amongst four scenarios. In addition, we have made point to point responses to all the ambiguous sentences with objective descriptions in the following technical comments raised by Referee #2.

B2S3: In addition, the authors tend to generalize in some cases findings that mainly apply to just two of the three catchments (see also my minor comments). I believe additional analyses may be needed to support the conclusions more, for example a statistical test to check if the distributions are significantly different. Or the addition of other, multiple performance measures, to assess the performance over multiple aspects (high flows, low flows etc.).

Reply: Thank you for your helpful comment. More information about the additional performance measures have been added in the revised manuscript.

(1) Firstly, two performance measures based on high flows (i.e., mean annual maximum flow) and low flows (i.e., mean annual minimum flow) have been used to evaluate the high and low flows. Please refer to lines 340-344, page 18.

(2) Secondly, the results of these measures have been be added as Tables 6 and 7 in the revised manuscript.

(3) Thirdly, discussions of the results of these measures have been added in lines 465-484, pages 24-25 in the revised manuscript.

B2S4: Further, all beta-values plot around zero in Figure 8, basically pointing at the absence of a clear trend. Is this indeed true? It would be interesting to show the timeseries of the parameter. The absence of a trend may explain the similar performances for all scenarios, and especially also why the time-varying scenarios do not outperform the others clearly. Besides, when beta is around zero, there is no point of looking at omega, as this does not do much in that case.

Concluding, the authors may need to clarify more what they did and how they arrive at several conclusions. I hope the authors find my comments useful, and I look forward to a revised manuscript.

Reply: We apologize for the mistakes in Figures 8 and 9.

(1) As discussed in response to comment A1 by Referee #1, Figures 8 and 9 have been redrawn.

(2) As discussed in the section 4.2, model parameter θ_1 is time-varying and no spatial coherence is considered ($\beta \neq 0$) in scenario 4, while θ_1 is stationary and of course no spatial coherence is included in scenario 5 (β=0). Scenario 4, had a higher median NSE$_{\mathrm{sqrt}}$ performance than that of scenario 5 in five of six options (except catchment 405219 in the first DSST scheme), which indicates the validity of the time-varying scheme for improving the model performance. Compared with scenario 5, the introduction of additional regression parameters (α, β and ω) in scenario 4 at the same time amplified the model projection uncertainty in two of three catchments (225219 and 405264). However, the appropriate adoption of spatial coherence alleviates this problem. Scenario 3, which considered both spatial coherence of regression parameters β and ω between different catchments, exhibited the optimal median NSE$_{\mathrm{sqrt}}$, DIC, and MaxF estimates in most options during the verification period, which illustrated the validity of the inclusion of the spatial coherence of regression parameters β and ω. Please refer to lines 389-454, pages 21-23.

(3) The median estimates of β in Figures 8 and 9 are not equal to zero. Because the facts that the adopted hydrological model is on a daily scale and the assumed time-varying model parameter θ_1 would change its values in each time step, the regression parameter β, as the amplitude of the sine term, is supposed to have a small absolute value rather a large one. As shown in Figure 8, the catchment average of the median estimate of β is 2.78 in scenario 1, -4.91 in scenario 2, 9.26 in scenario 3, and -39.20 in scenario 4 in the scheme of calibrating in the non-dry period

and verifying in the dry period. Please refer to lines 506-527, pages 26- 27.

(4) After calculation, parameter β has a deterministic posterior distribution (as shown in Figures 8 and 9) rather than a time-varying one.

Technical corrections

B3: P.7. section 2.1.1. Please elaborate on how the dry periods are defined.

Reply: Thank you. Please refer to response of comment A3 about the detailed definition of the dry period, which has been added in the revised manuscript.

B4: P8. Section 2.1.2. Why add this paragraph when you only refer to section 2.5?

Reply: This paragraph has been modified. Please refer to lines 180-183, page 9.

B5: P10. L210 Do you mean Eq. 1?

Reply: We are sorry for this oversight. The phase “Eq.2” has been revised as “Eq.1” in the revised manuscript.

B6: P10.L210 ...expected to the same. . . ! expected to be the same

Reply: Thanks. Change has been made as suggested.

B7: P12. L50. Please define N and n

Reply: Thanks. *N* refers to the Gaussian distribution and *n* represents the number of regression parameters that are spatially coherent. The definitions of *N* and *n* have been added in lines 287-288, page 14.

B8: P12.L258. Which parameters are optimized in SCEM-UA?

Reply: Thanks. All unknown quantities of different scenarios that needed to be optimized in SCEM-UA have been added in the supplementary material.

B9: P15.L326. Please explain how I can see this from Figure 4, except for the pre-defined red colour. Is this where the black line crosses the axis? Why are the first years not considered?

Reply: Thanks. The bars in blue and red colors in Figure 4 represent annual rainfall anomalies during the non-dry and dry periods, respectively. The black line is annual anomaly of rainfall smoothed with the 3-year moving window.

The start of the dry period is defined as the start of first 3-year consecutive negative anomaly period based on Saft et al (2015). According to the definition of dry period, the start of the dry period is not the place where the black line crosses the axis. Because the years near the cross point have positive rainfall anomaly. Similar comment is also raised by the Referee #1 (see comment A3 and our corresponding response). In the revised manuscript, the definition of the dry periods has been added. Please refer to lines 159-176, pages 8-9.

B10: P16.L339-340. Are these references in the right place? You describe your own results, shouldn't you refer to one of the figures?

Reply: Thank you for your comment. These references have been deleted, and the sentence has been modified as follows:

As shown in Figures 5(a), 6(a) and 7, the calibrated model parameters yielded good simulation performance over the calibrated periods for all criteria. Please refer to lines 389-390, page 20.

B11: P17.L355-357. This is, as far as I can see, not true for all catchments. Catchments 225219 and 405264 have a higher median, but the variation is less for 225219.

Reply: We apologize for this mistake. In figure 5(b), the variation of $\mathrm{NSE}_{\mathrm{sqrt}}$ in scenario 4 is less than that in scenario 5. The phrase of the comparison of variation has been deleted in the revised manuscript, because in this sentence we focus on the advantage of scenario 4, i.e., the improvement in median $\mathrm{NSE}_{\mathrm{sqrt}}$ performance. This sentence has been modified as follows:

Scenario 4 had a higher median $\mathrm{NSE}_{\mathrm{sqrt}}$ performance than scenario 5 in catchments 225219 and 405264, and was slightly inferior than the latter in catchment 405219, which indicates the validity of the time-varying scheme for improving the model performance. Please refer to lines 406-410, page 21.

B12: P17.L362. As far as I can see, it has only the highest median value for catchment 225219.

Reply: We apologize for our mistakes. These sentences have been modified as follows:

In the DSST scheme of calibrating in the dry period and verifying in the non-dry period, scenario 3, which both considered spatial coherence of regression parameters β and ω between different catchments, exhibited the highest median $\mathrm{NSE}_{\mathrm{sqrt}}$ for all catchments, had the smallest fluctuation range in two catchments (405219 and 405264) and is the second smallest scenario in catchment 22519 during the verification period. In the other DSST scheme, scenario 3 exhibited the smallest fluctuation range of $\mathrm{NSE}_{\mathrm{sqrt}}$ estimate for all catchments, showed the highest median value in catchment 225219, and was the second best scenario in the other two catchments (405219 and 405264) during the verification period. Please refer to lines 436-441, pages 22-23.

B13: P18.L375. The performances in the verification period seem higher to me? What do you mean calibrated performances were inferior?

Reply: We are sorry for the unclearness. In the scheme of calibration in dry period and verification in non-dry period, it is true that the $\mathrm{NSE}_{\mathrm{sqrt}}$ during the verification period is higher than that in the calibration period. However, the projection performance calibrated using a contrasting climatic condition was inferior to the simulation performance that was directly calibrated from the climatic condition, compared with Figures 5(a) and 6(b), or Figures 6(a) and 5(b). For example, the

NSE$_{sqrt}$ performance in Figure 6(b) is inferior to which in Figure 5(a). In the other words, for the non-dry period: $NSE_{sqrt,Figure\ 5\text{(a)}} > NSE_{sqrt,Figure\ 6\text{(b)}}$; for the dry period: $NSE_{sqrt,Figure\ 6\text{(a)}} > NSE_{sqrt,Figure\ 5\text{(b)}}$.

These sentences have been modified as follows in the revised manuscript:

"However, the projection performance calibrated using a contrasting climatic condition was inferior to the simulation performance that was directly calibrated from the climatic condition, compared with Figures 5(a) and 6(b), or Figures 6(a) and 5(b). For example, the NSE$_{sqrt}$ performance in Figure 6(b) is inferior to which in Figure 5(a)." Please refer to lines 427-431, page 22.

B14: P.18L375-377. This is not true for catchment 225219

Reply: We apologize for our mistakes. Follow the referee's comment, this sentence has been modified to specify the problem, which is as follows:

By comparing scenarios in the calibration period, it was found that scenarios 4 and 5 exhibited the highest performance in two of three catchments (405219 and 405264), followed successively by scenario 3, scenario 2, and scenario 1.

Please refer to lines 431-434, pages 22.

B15: P18.L379. The ranges seem not very different between scenarios 4 and 5, only slightly.

Reply: Thanks. These sentences have been modified as follows:

During the verification period, the median NSE$_{sqrt}$ performance in scenario 4 was 0.80% higher than scenario 5, however, the variation range in scenario 4 was 53% wider than the latter. In the DSST scheme of calibrating in the dry period and verifying in the non-dry period, scenario 3, which considered both spatial coherence of regression parameters β and ω between different catchments, exhibited the highest median NSE$_{sqrt}$ for all catchments, had the smallest fluctuation range in two catchments (405219 and 405264) and is the second smallest scenario in catchment 22519 during the verification period.

Please to lines 434-441, pages 22-23.

B16: P18.L379-380. It's not very obvious that scenario 3 has a higher median performance for catchment 405264.

Reply: Thanks. This sentence has been deleted.

B17: P18.L382 This is not very obvious to me.

Reply: Thank you. These sentences have been modified. Please refer to response to comment 15.

B18: P18.L394. Compared –> comparing

Reply: Thanks. Change has been made as suggested.

B19: P20.L438. Is omega for scenario 4 not the lowest in all cases? Or do you mean the absolute values?

Reply: We apologize for our mistakes. This should be regression parameter β in this place rather than ω. The catchment averages of the median estimates of β in the first three scenarios are 2.78, -4.91, and 9.26 respectively, while that in the fourth scenario is much larger, reached at -39.20. Scenario 3, which considers both spatial coherence of regression parameters β and ω, has the narrowest interval of β for all catchments, followed successively by scenario 1 (only considered the spatial coherence of the regression parameter β), scenario 2 (only regression parameter ω was spatially coherent), and scenario 4 (no regression parameter was spatially coherent). With regards to the regression parameter ω, which denotes the frequency of the sine function (in the lower figures of Figures 8 and 9), its median estimates and variation ranges in both four scenarios differ slightly. The former reached a catchment average of 0.19,0.20,0.19,0.17 for different scenarios.

B20: Figure 2. Please define all symbols and abbreviations in the figure.

Reply: Thanks. All symbols and abbreviations in the Figure 2 have been defined in the revised manuscript.

B21: Figure 5,6: I would suggest to plot the boxes for calibration and verification next to each other. It's easier to see whether there is an improvement or not. Please also add the units (also when a unitless number is presented)

Reply: Thanks. Changes have been made as suggested.

B22: Figure 7. Please make the labels and text bigger.

Reply: Thanks. Changes have been made as suggested.

B23: Figure 8, 9. Maybe use the same colors for the scenarios in both plots. What are the units of beta and omega?

Reply: Thanks for thoughtful comment. Similar comment has been raised by Referee #1 (see comment A1). The same color for the scenarios in both plots has been used. Both β and ω are dimensionless. Figures 8 and 9 have been modified as violin plots in the revised manuscript (see response to A1).