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: We apologize for this oversight.

(1) All the hyper-parameters are not determined in advance. The hyper-parameters are sampled and determined with other unknown quantities simultaneously 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 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, the final parameter samples that satisfy the requirement that a GR value must be smaller than a Gelman-Rubin convergence value of 1.2 (Gelman et al., 2013) would be selected as the posterior probability distribution of parameters.

(2) The ω 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 would be sampled and derived as independent variables. For example, in scenario 4, regression parameters ω and β of different catchments are not linked, thus values of ω and β of each catchment are calibrated from corresponding catchment inputs. In scenario 1, regression parameter $\beta(c) = N(\mu_3, \sigma^2)$, which means that β is shared with linked catchments, while independent regression parameters $\omega_{1\text{-}1}$, $\omega_{1\text{-}2}$, and $\omega_{1\text{-}3}$ are used to represent the frequency of model parameter θ_1 in different catchments. The name of all unknown quantities in different scenarios could be found 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.

(1) The Gaussian distribution is one of widely used distributions for describing the process level 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 literatures.

(2) Only the structure of Gaussian distribution and ranges of all unknown quantities were fixed in advance. The hyper-parameters and the prior distributions were obtained from Sun et al (2015,2016). All unknown quantities would be derived in the SCEM-UA algorithm. In addition, as illustrated in response to comment A14 by Referee #1, convergence for the SCEM-UA algorithm is assessed by evolving three parallel chains with 30000 random samples, combined with the additional large prior ranges for all unknown quantities, is enough for ensuring a reliable result of all parameters.

(3) The prior ranges of all unknown quantities in different scenarios are 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.

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.

(1) The homogeneity assumption will 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 Figure 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 will be 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) will be used to evaluate the high and low flows. The following paragraph will be added in section 2.5 in the revised manuscript.

The fourth and fifth criteria are the Mean annual maximum flow (MaxF, mm/d) and Mean annual minimum flow (MinF, mm/d), which are used to qualify the performance of the high flows and low flows. These criteria are self-explanatory and have been used in many studies to assess the magnitude of maximum and minimum levels of flows (Ekstrom et al., 2018).

(1) Secondly, the following paragraph about the results of these measures will be added as Tables 6 and 7 in the revised manuscript.

Table 6. Comparison of the projection performance during the verification period regarding the mean annual maximum flow (MaxF, mm/d) and mean annual minimum flow (MinF, mm/d) when model parameters were calibrated in the non-dry period and verified in the dry period.

	Mean annual maximum flow			Mean annual minimum flow		
	225219	405219	405264	225219	405219	405264
Observed	10.58	11.98	9.23	0.050	0.093	0.17
Scenario 1	13.30	5.64	6.68	**0.050**	0.045	0.13
Scenario 2	9.04	**10.23**	7.30	0.054	0.060	0.14
Scenario 3	**10.91**	7.66	9.75	0.041	**0.092**	**0.16**
Scenario 4	5.91	5.42	**9.54**	0.089	0.089	0.15
Scenario 5	5.07	6.03	7.98	0.086	0.086	0.12

Note: 1. The data in 1976 has been used for model warm-up to reduce the impact of the initial soil moisture conditions during the calibration period, and is not counted in the table; 2. The scenarios with bold values are labeled as the best scenario for projecting the streamflow during the verification periods.

Table 7. Comparison of the projection performance during the verification period associated with the Mean annual maximum flow (MaxF, mm/d) and Mean annual maximum flow (MinF, mm/d) when model parameters were calibrated in the dry period and verified in the non-dry period.

	Mean annual maximum flow	Mean annual minimum flow

	225219	405219	405264	225219	405219	405264
Observed	10.73	12.06	8.94	0.03	0.09	0.19
Scenario 1	12.40	6.87	12.90	0.03	0.04	0.09
Scenario 2	12.42	5.52	10.30	0.02	**0.06**	0.09
Scenario 3	**10.95**	**10.67**	**8.37**	**0.03**	0.05	0.10
Scenario 4	11.98	9.85	12.34	0.03	0.05	**0.10**
Scenario 5	14.19	9.45	11.97	0.02	0.05	0.10

Note: 1. The data in 1997 has been used for model warm-up to reduce the impact of the initial soil moisture conditions during the calibration period, and is not counted in the table; 2. The scenarios with bold values are labeled as the best scenario for projecting the streamflow during the verification periods.

(2) Thirdly, discussions of the results of these measures will be added in P21-22. L445-463 in the revised manuscript, which are as follows.

Tables 6 and 7 illustrate the performance of high and low flows during the verification period in terms of MaxF and MinF estimates for the median projected streamflows in both DSST schemes. As shown in table 7, for the projection of high flow part, scenario 3 exhibits the best performance in all catchments among five scenarios under the scheme of calibrating in the dry period and verifying in the non-dry period. For the projection performance in the other DSST scheme (Table 6), scenario 3 has the best projection performance in high flow part in catchment 225219 and is the second best scenario in the other two catchments. It indicates that the incorporation of spatial coherence of both regression parameters β and ω successfully improves the projection performance in the high flow part. As for the projection of the low flow part, the discrepancy between the results of different scenarios and the observed low flows is not obvious. Furthermore, scenario 3 shows the best projected performance in two catchments (405219 and 405264) in the scheme of calibrating in dry period and verifying in non-dry period, and is the best scenario in catchment 405264 in the scheme

of calibrating in non-dry period and verifying in dry period. In addition, scenario 3 is the second best option in catchment 225219 and 405219 under the scheme of calibrating in non-dry period and verifying in dry period. Combined with the projection performance of both high and low flows, scenario 3 achieves its superior projection performance mainly by the improvement in the prediction of high flow parts.

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 Figure 8 and 9.

(1) As response to comment A1 by Referee #1, Figures 8 and 9 will be redrawn as follows:

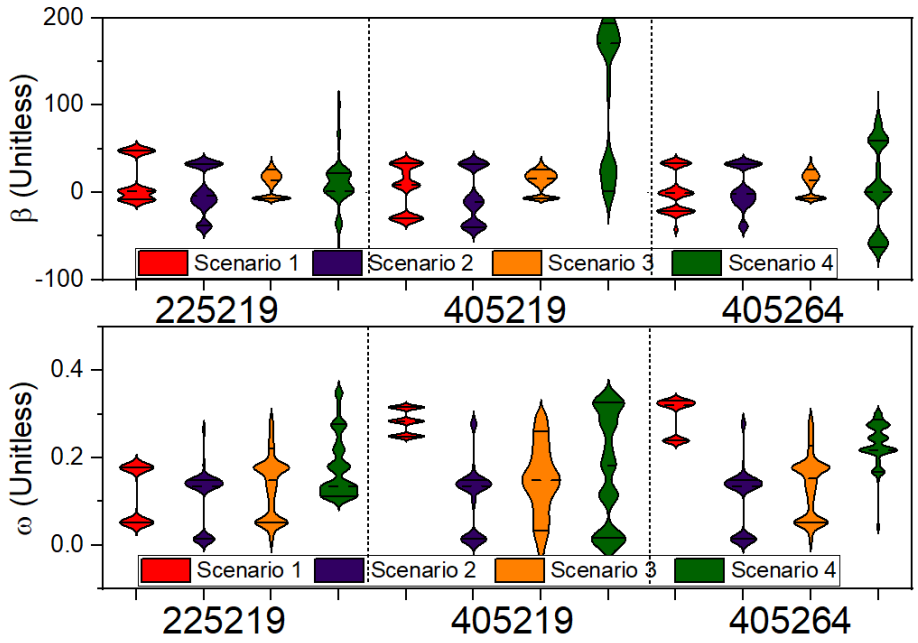

Figure 8. Posterior distributions of the regression parameters (β and ω) for the production storage capacity (θ_1) for the four model scenarios in each catchment

when calibrated in the non-dry period and 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 dotted line denotes median estimates.

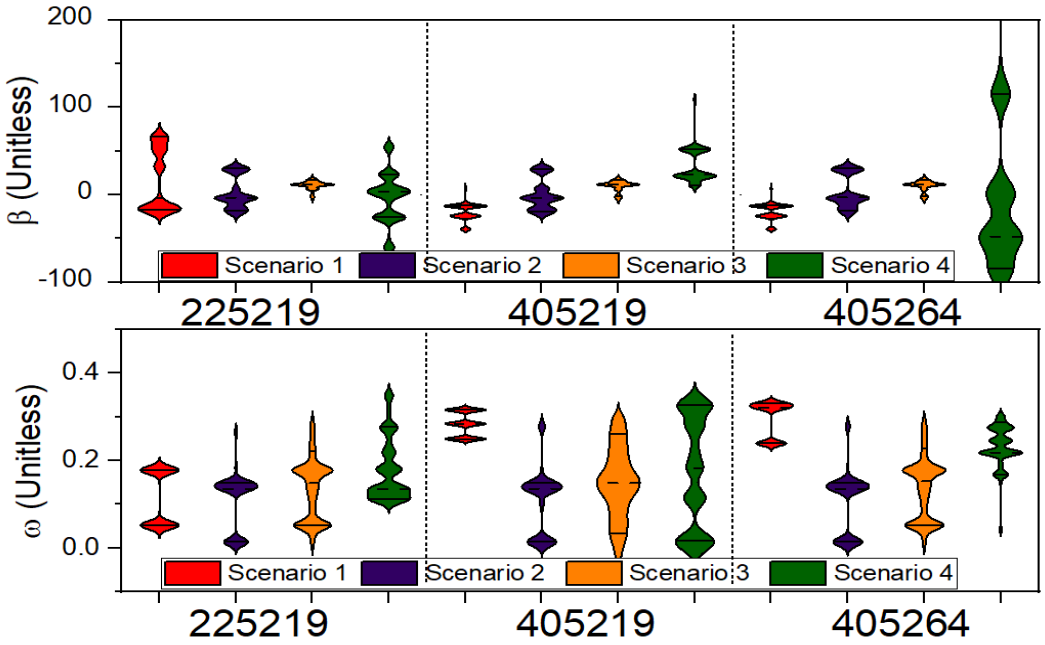

Figure 9. Posterior distributions of the regression parameters (β and ω) for the production storage capacity (θ_1) for the four model scenarios in each catchment when calibrated in the dry period and verified in the non-dry period. The solid horizontal lines within the violin plots denote the 25th and 75th percentiles of the posterior distribution, while the dotted line denotes median estimates.

(2) As discussed in the section 3.2.2, 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_{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 both considered spatial coherence of regression parameters

β and ω between different catchments, exhibited the optimal median NSE_{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.

(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 are 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.

(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 will be added in the revised manuscript.

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

Reply: We apologize for this misunderstanding. This paragraph will be modified as follows:

In the DSST method, the model parameters calibrated in the non-dry period were evaluated in the dry period, and vice versa. In addition, criteria, i.e, NSE_{sqrt}, BIAS and

DIC illustrated in the section 2.5, were used to evaluated the performance of the calibrated parameters for different transfer schemes.

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

Reply: We are sorry for this oversight. The phase "Eq.2" will be revised as "Eq.1" in the revised manuscript.

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

Reply: Thanks. Change will be 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* will be added in the revised manuscript.

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 will be added.

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 will be deleted, and the sentence will be 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.

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 NSE_{sqrt} in scenario 4 is less than that in scenario 5. The phrase of the comparison of variation will be deleted in the revised manuscript, because in this sentence we focus on the advantage of scenario 4, i.e., the improvement in median NSE_{sqrt} performance. This sentence will be modified as follows:

Scenario 4 had a higher median NSE_{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.

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. This sentence will be 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 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. In the other DSST scheme, scenario 3 exhibited the smallest fluctuation range of NSE_{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.

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 this misunderstanding. In the scheme of calibration in dry period and verification in non-dry period, it is true that the NSE_{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 Figure 5(a) and 6(b), or Figure 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: $NSE_{sqrt,Figure\ 5(a)} > NSE_{sqrt,Figure\ 6(b)}$; for the dry period: $NSE_{sqrt,Figure\ 6(a)} > NSE_{sqrt,Figure\ 5(b)}$.

This sentence will be 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 Figure 5(a) and 6(b), or Figure 6(a) and 5(b). For

example, the NSE_{sqrt} performance in Figure 6(b) is inferior to which in Figure 5(a)." Hope the revision is clear to the referee and readers.

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

Reply: We apologize for our mistakes. Follow the referee's comment, this sentence will be 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.

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

Reply: Thanks. This sentence will be 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 both considered 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.

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

Reply: Thanks. This sentence will be modified as follows:

In catchment 405264, compared with scenarios 1, 2, 4 and 5, scenario 3 showed

an 8.1%, 6.7%, 0.2% and 0.6% improvement in the median NSE_{sqrt} performance, respectively.

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

Reply: Thank you. This sentence will be 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 both considered 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.

B18: P18.L394. Compared –> comparing

Reply: Thanks. Change will be 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 average 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 considered 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 parameter ω was spatially

coherent), and scenario 4 (no 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 will be defined in the revised manuscript.

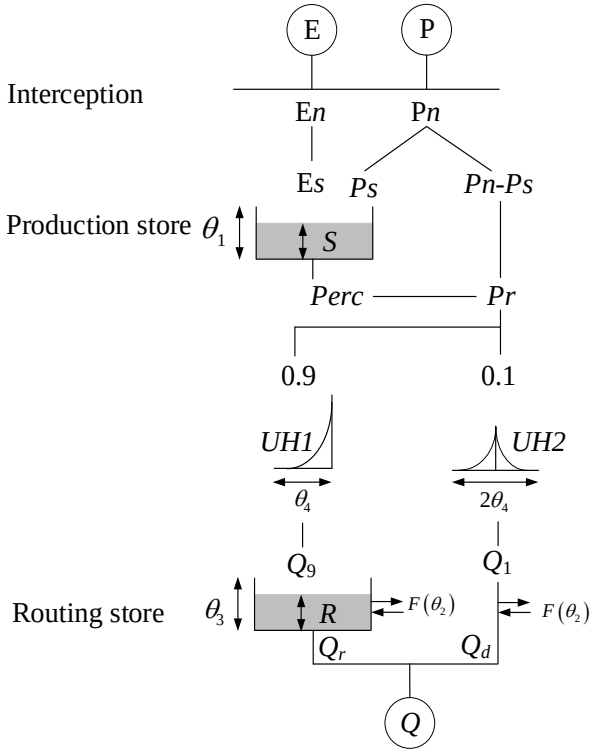

Figure 2. Schematic diagram of the GR4J rainfall-runoff model adopted from Perrin et al. (2003). In the figure, P and E refer to precipitation and evapotranspiration, respectively; E_n and P_n denote net precipitation and net evapotranspiration, respectively; Ps refers part of precipitation that fills the production store (i.e. S). The production store is determined as a function of the water level S in production store. The θ_1,θ_2,θ_3, and θ_4 denote model parameters. The *Perc* refers to the percolation leakage that is a function of production store *S* and parameter θ_1. The Pr refers to total quantity of water that reaches the routing functions. The UH1 and UH2 denote two unit hydrographs. The Q_1 and Q_9 refer the corresponding output of the unit hydrographs, respectively; F indicates

the groundwater exchange term; R is the level in the routing store. The Q_r refers to the outflow of the routing store, Q_d is a function of water exchange, and Q refers to the total streamflow.

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 will be made as suggested.

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

Reply: Thanks. Changes will be made as suggested. The modified Figure 7 will be as follows:

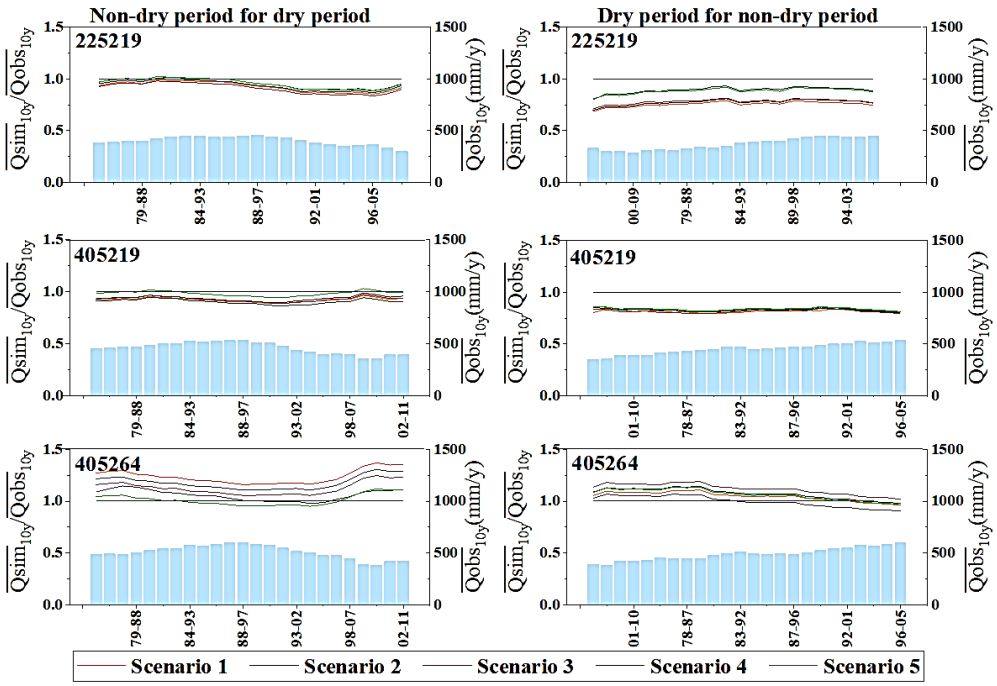

Figure 7. BIAS performance of Q_{median} for five scenarios in all catchments. The BIAS is plotted as a 10-year moving average, and 10-year moving average streamflows are plotted for reference. The left-hand three graphs are calibrated in the non-dry period and then verified in the dry period, while the opposite sequence applies to the right-hand graphs.

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 will be used. Both

β and ω are unitless. Figures 8 and 9 will be modified as violin plots in the revised manuscript (see response to A1).

Supplement:

Table S1 The prior ranges of all unknown quantities in different scenarios

(1) Calibration in non-dry period and verification in dry period:

Scenario 1:

$\theta_{2\text{-}1}$	$\theta_{2\text{-}1}$	$\theta_{2\text{-}3}$	μ_2	σ_3	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\omega_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\omega_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\omega_{1\text{-}3}$
-10	-10	-10	-100	0	0.1	1	1	0.0001	0.1	0.5	100	0.0001	0.1	0.1	1	0.0001
10	10	10	100	6	200	10	600	0.4	300	20	1000	0.4	300	20	500	0.4

Scenario 2:

$\theta_{2\text{-}1}$	$\theta_{2\text{-}1}$	$\theta_{2\text{-}3}$	μ_3	σ_3	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\beta_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\beta_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\beta_{1\text{-}3}$
-6	-6	-6	-0.4	0	1	0.5	1	-300	1	0.1	100	-300	0.1	2	1	-200
-6	-6	-6	0.4	0.1	500	10	600	300	300	20	600	500	400	20	800	300

Scenario 3:

$\theta_{2\text{-}1}$	$\theta_{2\text{-}1}$	$\theta_{2\text{-}3}$	μ_2	σ_2	μ_3	σ_3	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$
-5	-5	-5	-200	0	-0	0	1	0.5	1	1	0.1	100	1	0.5	100
5	5	5	100	8	0.4	0.1	120	10	500	300	20	500	250	20	600

Scenario 4:

$\theta_{2\text{-}1}$	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\beta_{1\text{-}1}$	$\omega_{1\text{-}1}$	$\theta_{2\text{-}2}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\beta_{1\text{-}2}$	$\omega_{1\text{-}2}$	$\theta_{2\text{-}3}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\beta_{1\text{-}3}$	$\omega_{1\text{-}3}$
-10	1	0.1	1	-300	0.0001	-10	1	0.1	0	-300	0	-10	1	0.1	0	-300	0.0001
10	500	10	800	300	0.4	10	500	10	800	300	0.4	10	500	10	800	300	0.4

(2) Calibration in dry period and verification in dry period:

Scenario 1**:**

$\theta_{2\text{-}1}$	$\theta_{2\text{-}2}$	$\theta_{2\text{-}3}$	μ_2	σ_2	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\omega_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\omega_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\omega_{1\text{-}3}$
-10	-10	-10	-60	0	1	0.5	1	0	1	0.5	1	0	1	0.1	1	0
10	10	10	60	6	300	10	600	0.4	300	20	600	0.4	300	15	600	0.4

Scenario 2:

$\theta_{2\text{-}1}$	$\theta_{2\text{-}2}$	$\theta_{2\text{-}3}$	μ_3	σ_3	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\beta_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\beta_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\beta_{1\text{-}3}$
-10	-10	-10	0.0001	0	1	0.5	1	-300	1	0.1	1	-400	0.1	0.5	1	-400
10	10	10	0.4	0.1	200	15	500	400	300	20	600	500	140	20	600	400

Scenario 3:

$\theta_{2\text{-}1}$	$\theta_{2\text{-}2}$	$\theta_{2\text{-}3}$	μ_2	σ_2	μ_3	σ_3	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$
-10	-10	-10	-80	0	0	0	1	0.5	1	1	0.1	1	1	0.1	1
10	10	10	80	6	0	0.1	200	10	500	400	20	600	400	20	600

Scenario 4:

$\theta_{2\text{-}1}$	$\theta_{3\text{-}1}$	$\theta_{4\text{-}1}$	$\alpha_{1\text{-}1}$	$\beta_{1\text{-}1}$	$\omega_{1\text{-}1}$	$\theta_{2\text{-}2}$	$\theta_{3\text{-}2}$	$\theta_{4\text{-}2}$	$\alpha_{1\text{-}2}$	$\beta_{1\text{-}2}$	$\omega_{1\text{-}2}$	$\theta_{2\text{-}3}$	$\theta_{3\text{-}3}$	$\theta_{4\text{-}3}$	$\alpha_{1\text{-}3}$	$\beta_{1\text{-}3}$	$\omega_{1\text{-}3}$
-10	1	0.1	1	-300	0.0001	-10	1	0.1	1	-300	0	-10	1	0.1	1	-300	0
10	500	10	800	300	0.4	10	500	10	800	300	0.4	10	500	10	800	300	0.4

Notes:

$\theta_{2\text{-}1}$, $\theta_{2\text{-}2}$ and $\theta_{2\text{-}3}$ refers to model parameter θ_2 in catchment 225219, 405219 and 405264, respectively; $\theta_{3\text{-}1}$, $\theta_{3\text{-}2}$ and $\theta_{3\text{-}3}$ refer to model parameter θ_3 in catchment 225219, 405219 and 405264, respectively; $\theta_{4\text{-}1}$, $\theta_{4\text{-}2}$ and $\theta_{4\text{-}3}$ refers to model parameter θ_4 in catchment 225219, 405219 and 405264, respectively; μ_2, σ_2, μ_3 and σ_3 represent four hyper-parameters; $\alpha_{1\text{-}1}$, $\alpha_{1\text{-}2}$ and $\alpha_{1\text{-}3}$ refer to regression parameter α in catchment 225219, 405219 and

405264, respectively; $\beta_{1\text{-}1}$, $\beta_{1\text{-}2}$ and $\beta_{1\text{-}3}$ refer to regression parameter β in catchment 225219, 405219 and 405264, respectively; $\omega_{1\text{-}1}$, $\omega_{1\text{-}2}$ and $\omega_{1\text{-}3}$ refer to regression parameter ω in catchment 225219, 405219 and 405264, respectively.

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

$$\textit{Scenario } 1\text{: } \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta,\omega(c)\right]\bullet f_{N}\left(\beta\middle|\mu_2,\sigma_2\right)$$

$$\textit{Scenario } 2\text{: } \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta(c),\omega\right]\bullet f_{N}\left(\omega\middle|\mu_3,\sigma_3\right)$$

$$\textit{Scenario } 3\text{: } \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\middle|\alpha(c),\beta,\omega\right]\bullet\prod_{n=1}^{2} f_{N}\left(\beta,\omega\middle|\mu_2,\sigma_2,\mu_3,\sigma_3\right)$$

$$\textit{Scenario } 4\text{: } \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(t,c),\theta_2(c),\theta_3(c),\theta_4(c)\right]$$

$$\textit{Scenario } 5\text{: } \Lambda=\prod_{c=1}^{C}\varepsilon_c\left[\theta_1(c),\theta_2(c),\theta_3(c),\theta_4(c)\right]$$