Journal:	Hydrology and Earth System Sciences
Title:	Multivariate calibration can increase simulated discharge uncertainty and model equifinality
Manuscript:	egusphere-2025-1598
Authors:	Sandra Pool, Keirnan Fowler, Hansini Gardiya Weligamage, and Murray Peel

Dear Julia Knapp, dear reviewers,
We thank you for your efforts with our manuscript and the very constructive and valuable feedback. Please see the point-by-point response to the two reviewers' comments below. We also made a few other changes, which we indicated in the revised manuscript with the tracked changes.
Kind regards, also on behalf of all co-authors,
Sandra Pool

Reviewer 1 (Tam Nguyen)

Summary

The study by Pool et al., titled "Multivariate calibration can increase simulated discharge uncertainty and model equifinality", investigates the uncertainty associated with simulated discharge and the flux components (overland flow, interflow, and baseflow) under both univariate and multivariate calibration approaches. In the univariate calibration, the hydrological model was calibrated either for discharge (Q) or for actual evapotranspiration (ET), whereas in the multivariate calibration, the model was simultaneously calibrated for both Q and ET using a combined objective function. The analysis was conducted across twelve Australian catchments utilizing the SIMHYD hydrological model. The authors report that multivariate calibration—through the additional inclusion of ET—leads to increased uncertainty in both simulated discharge and the associated flux maps. This is attributed to the absence of "overlapping behavioral parameter sets" (Fig. 4).

Overall, the research demonstrates a high degree of novelty, as it is likely the first study to explore this topic in such depth. The manuscript is well written, the results convincingly support the conclusions, and the discussion is comprehensive.

Major comments

Comment 1: In the manuscript, the authors refer to model structure at several points (e.g., lines 23–25, 117–119, and lines 79–80: "A lack of intersection might be a symptom of problems such as an inadequate model structure … or data that are subject to systematic biases or other error"). In this context - specifically regarding the lack of intersection in behavioral parameter sets shown in Figure 4 - I am curious whether this outcome is due to limitations in the SIMHYD model structure or issues with the input data. Furthermore, presenting scatter plots of the objective function values for Q and ET from the 100,000 simulations for each basin would be insightful. Such visualizations could help identify cases of trade-offs (negative correlation) or cross-benefits (positive correlation) between the two objective functions, and potentially clarify whether these patterns arise from model structural limitations or data-related issues.

Reply: The question of whether the lack of overlap stems from model structural inadequacy or data quality issues is indeed a very interesting and relevant one. Our analysis was conducted for two different ETa datasets (one remote sensing-based and one field-based) and led to very similar conclusions. We therefore assumed that ETa data-related issues are not the main reason for our observations (see discussion L389-L405).

So far, our analysis was largely based on flux maps and parameter distributions, and we followed the suggestion of the reviewer to also consider model performance. As suggested, we created scatter plots with the objective function values for discharge O_Q and actual evapotranspiration O_{ETa} from the

100,000 simulations (Figs. 1 and 2 in this response document or Figs. S4 and S8 in the supporting information of the revised manuscript). Correlations between the two objective functions O_Q and O_{ETa} are for the majority of catchments negative, suggesting a trade-off between the two objective functions. The sign of the rank correlation coefficient is, with the exception of two catchments, the same for both ETa datasets despite the very different nature of the two ETa datasets. This supports our current interpretations that data limitations alone cannot explain the observed trade-off and model structure could be the reason for the increased equifinality and discharge uncertainty under multivariate calibration. In the revised manuscript, we report the negative correlations in the results and the discussion sections, and we show the scatterplots in the supporting information.

Modification: L264-265, L400-401, Figs. S4 and S8

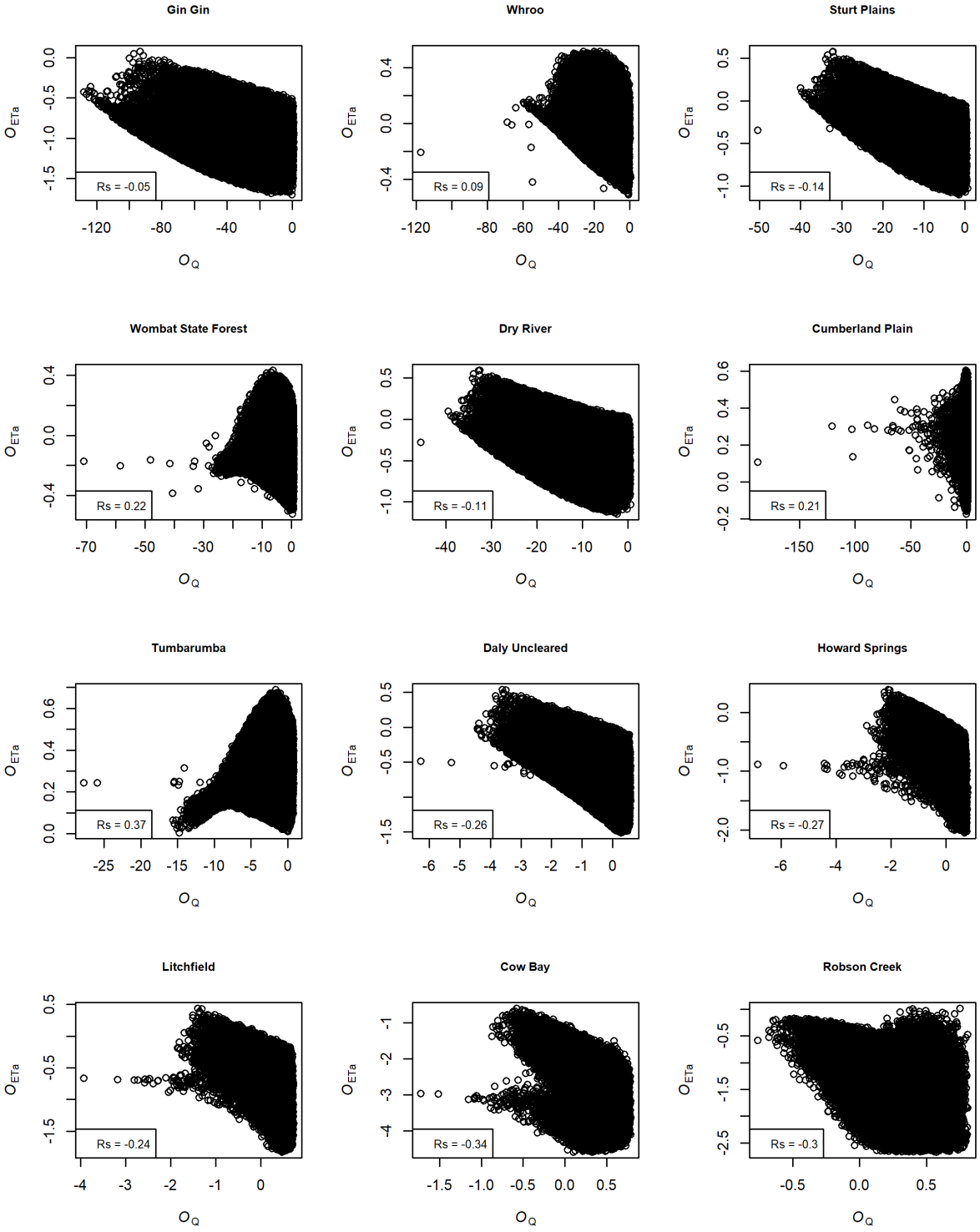

Figure 1: Model performance for discharge (O_Q) and actual evapotranspiration (O_{ETa}) for all 100,000 randomly selected parameter values. Rs values show the Spearman rank correlation coefficients between the two objective functions. Results are shown for the dataset using remote sensing-based actual evapotranspiration ($D_{ETa.RS}$).

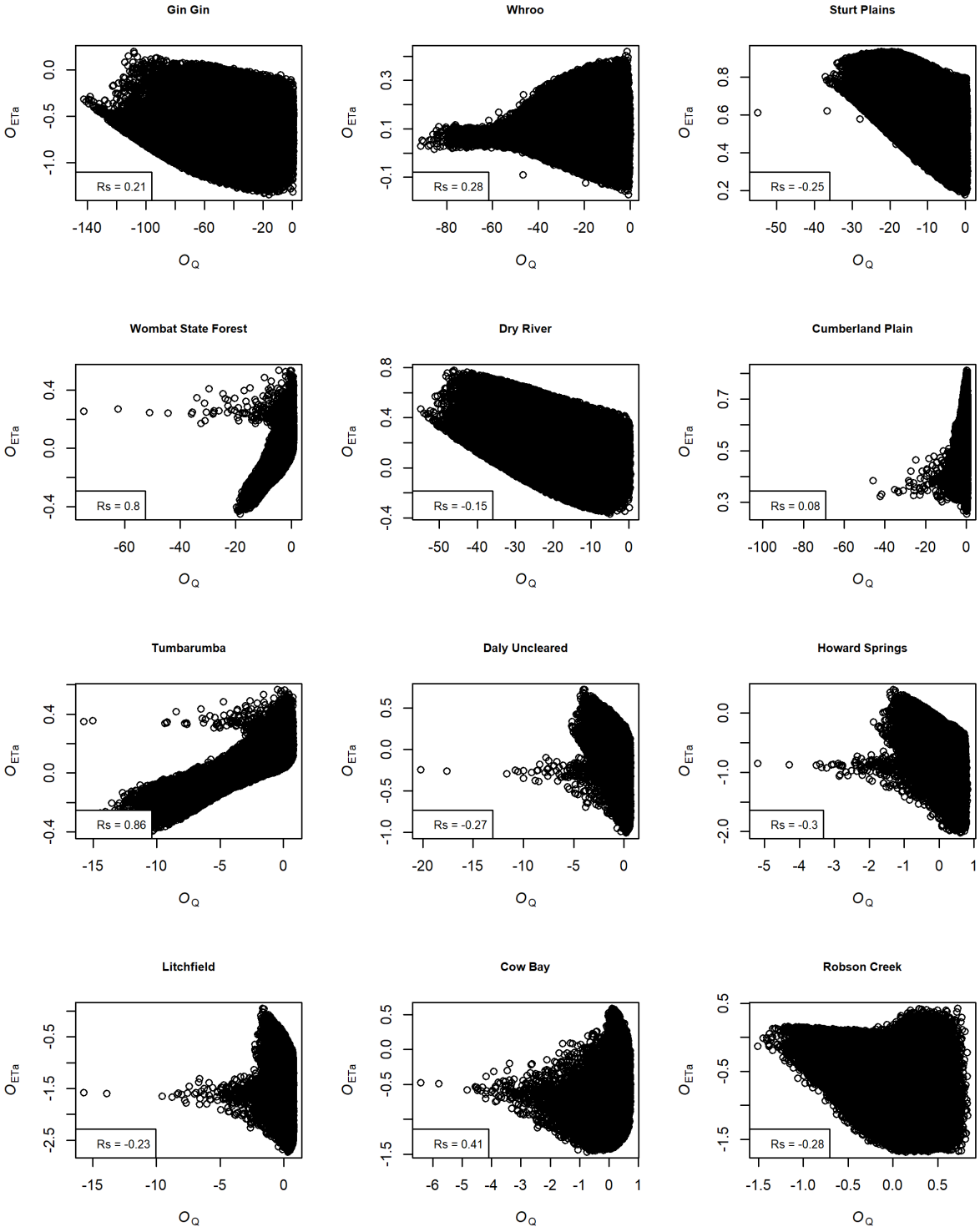

Figure 2: Model performance for discharge (O_Q) and actual evapotranspiration (O_{ETa}) for all 100,000 randomly selected parameter values. Rs values show the Spearman rank correlation coefficients between the two objective functions. Results are shown for the dataset using field-based actual evapotranspiration ($D_{ETa.Field}$).

Minor comments

Comment 2: In this study, the authors employed a combined objective function approach. A potential drawback of this method is that model performance may become biased toward one variable at the expense of another. This trade-off can contribute to increased uncertainty, as improvements in the

simulation of one variable may be offset by declines in the performance of another, and therefore, increase the uncertainty under multivariate calibration.

Reply: We agree with the reviewer that the observed increased uncertainty likely arises from the trade-off between the two calibration variables. The scatterplots presented in the response to comment 1 show that trade-offs (i.e., negative correlation between the two calibration variables) are present in most catchments and for both ETa datasets. Further examples are provided in Figs. S2 and S6 of the revised manuscript that contain boxplots of model performance for discharge and ETa for all calibration cases and catchments. They indicate that trade-offs exist for both variables, i.e., model performance for discharge decreases if adding ETa to a discharge-based calibration and vice versa. Trade-offs in model performance have been observed in various studies irrespective of the calibration approach. An important difference to other studies is that this trade-off leads to increased simulated discharge uncertainty, which we can attribute to non-overlapping behavioural parameter clouds. Our work, therefore, asks the critical question of when does the trade-off become so large that the addition of data doesn't help. We mention the limitation of the combined objective function approach and its impacts on our results in the discussion of the revised manuscript.

Modification: L413-419

Comment 3: Line 54-65. Did these studies employ separate objective functions for each variable in their multivariable calibration approach?

Reply: All three studies use separate objective functions for each variable. The difference is that Hartman et al. (2017) evaluate multiple variables simultaneously, whereas Arciniega-Esparza et al. (2022) and Kelleher et al. (2017) apply the calibration variables sequentially. We will adapt the text to provide this information.

Modification: 56-59

Comment 4: Eq. (2) please write this equation in full form?

Reply: We changed the equation as suggested to better highlight the modification applied to KGE.

Modification: L220-221, equation 2

Comment 5: Figure 4: Does the axis show the normalized values [0,1] instead of the actual values? If yes, please specify in the text. Adding x and y values to the subplots (b) even their ranges are [0, 1]?

Reply: Yes, we normalized the values of all model parameters in the scatterplots showing parameter values. We updated the figure caption accordingly and added the values to the subplots. To be consistent, we will also added the values to the subplots in the flux maps.

Modification: Figs. 5 and 7, Figs. S11-S12, including figure captions

Reviewer 2

Summary

I reviewed the manuscript entitled “Multivariate calibration can increase simulated discharge uncertainty and model equifinality” by Pool et al. This study employs a Monte Carlo approach (100,000 parameter sets) to compare three hydrological model calibration strategies: discharge-only, ET-only, and discharge+ET. The key finding—that incorporating ET into calibration increases flux uncertainty and discharge variability while revealing incompatible parameter spaces for discharge and ET objectives—is compelling and well-supported. However, two critical issues require resolution before publication (see below). There are also some specific comments in the annotated manuscript. Overall, I would suggest a major revision.

Major Concern: Definition of "Behavioral Parameters"

Comment 1: The authors defined the top 100 parameter sets as “behavioral parameters” based solely on their ranking without explicitly addressing the performance thresholds associated with these sets. For example: What NSE/KGE values correspond to the top 100 sets? Does the performance of the 100th set differ significantly from the 101st? Could expanding the threshold to the top 1,000 sets yield comparable or better performance while reducing equifinality? Since the “behavioral” classification directly shapes the conclusions (e.g., parameter space in compatibility in Section5.2), this ambiguity undermines the practical relevance of the results. Therefore, I would suggest to add performance metrics (e.g., NSE/KCE ranges) for the top 100/1,000 parameter sets in a table or figure and to justify the choice of 100 sets (e.g., via sensitivity tests or breakpoint analysis of performance vs. parameter set rank).

Reply: We acknowledge that the selection of the top 100 parameter sets as behavioural is a subjective choice, which can affect the conclusions. Following the suggestion of the reviewer, we created a figure showing model performance and parameter values for different thresholds of behavioural parameter sets (Fig. 3 in this response document or Fig. 8 in the revised manuscript). The analysis suggests that loosening the threshold (e.g., from 100 to 1000 behavioural parameter sets) may increase the chance of finding overlapping parameter distributions for univariate calibrations with discharge and actual evapotranspiration. However, this would come at the expense of accepting lower model performance values for the calibration variables although the extent of these costs differs among catchments and the threshold (Fig. 3a in this response document). The choice of a threshold also affects how much parameter values are constrained. Results for the parameter values are more consistent among catchments than for model performance and generally indicate that parameter values get quickly less constrained with an increasing number of behavioural parameter sets (Fig. 3b in this response document). Most parameters in SIMHYD directly affect runoff generation and a wider range of their values can further increase flux equifinality and thus hydrograph uncertainty. This is similar to findings reported in a study by Lamb et al. (1998), where less constrained TOPMODEL parameter values and wider discharge uncertainty bounds were found for multivariate calibration than in a discharge-based calibration. Thus, working with the best 100 parameter sets seems to be a reasonable decision for this study.

Modification: L423-430, Fig. 8

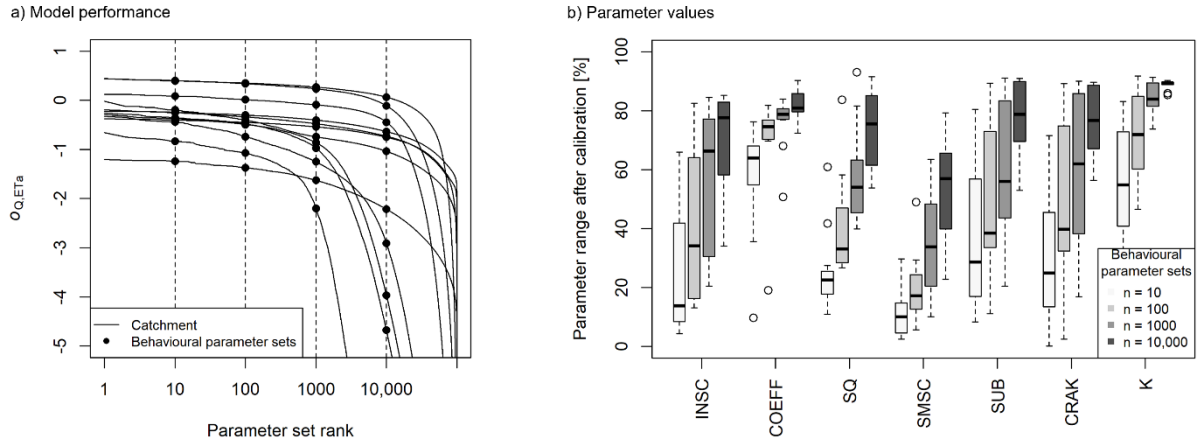

Figure 3. Definition of behavioural parameter sets and its impact on a) model performance $O_{Q,Eta}$ and b) parameter values, i.e., how much a parameter is constrained after calibration with $O_{Q,Eta}$ relative to the range of possible parameter values before calibration (see Fig. S1 for an explanation of the parameters). Results are shown for all twelve study catchments.

Moderate Concern: Parameter Distribution Analysis (Figure 4)

Comment 2: Figure 4 illustrates parameter distribution overlaps between calibration strategies but lacks methodological transparency: The rationale for selecting specific parameter pairs (e.g., sensitivity, correlation) is unclear. Subjective terms like "separation" need quantitative support (e.g., Kolmogorov-Smirnov tests or overlap coefficients). I would suggest to clarify how parameter pairs were chosen and to replace qualitative descriptions with metrics (e.g., heatmaps of distribution overlap or pairwise correlation matrices) to synthesize relationships across all parameters.

Reply: As correctly pointed out by the reviewer, we didn't use a quantitative metric for measuring the lack of overlap between parameter distributions. This is because, in the two-dimensional space, there are numerous examples for which the lack of overlap is visually very clear. Having these examples was enough to reject our hypothesis. However, we agree with the reviewer that a quantitative metric for the overlap of parameter value distributions can be beneficial, especially to present results for all possible parameter combinations. We therefore computed the overlap of the convex hull of the behavioural parameter clouds for each parameter pair and summarized the results for each catchment in a heatmap (Fig. 4 of this response document or Fig. 6 in the revised manuscript).

The parameter pairs demonstrating the lack of overlap in the scatterplot in Fig. 6 were manually selected to show different separation patterns. We adapted the sentence to be more clear about the rational of our choice.

Modification: L323-328, Fig. 6

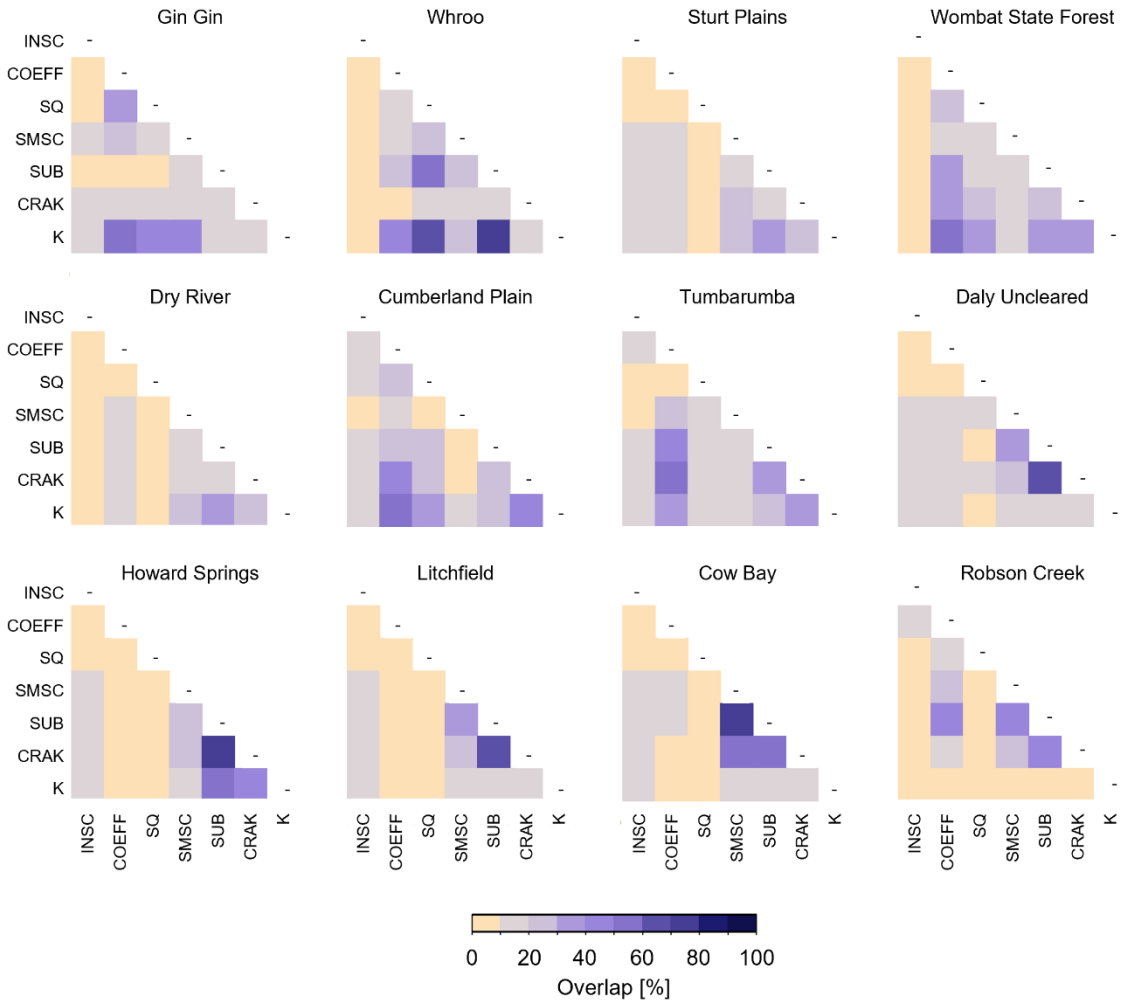

Figure 4. Overlap of the two point clouds of behavioural parameter sets for a given parameter pair when calibrating the model solely to discharge (O_Q) and solely to actual evapotranspiration (O_{ETa}). The orange squares show the parameter combinations with a lack of overlap in the two-dimensional space.

Specific comments

L45: The 20 catchments used in this study are all within the United States.

Reply: Thanks for pointing this out. We adapted the sentence to refer to the correct region.

Modification: L45

L115: This is an important analysis technique but may not be familiar to everyone. I would suggest to add a couple of sentences (without technical details and directing the authors to section 2.3.3) some where from line 95 to 120.

Reply: We added some more information on the idea of Flux Mapping where we first introduce the concept.

Modification: L96-102

L122: Delete duplicate title.

Reply: We removed the duplicate title.

Modification: L129

L170: I think it is complicated to have sub-order for panel index. I would suggest to call these b, c, d.

Reply: We adapted the labels of the sub-panels to a-d.

Modification: Caption of Fig 1.

L201: (KGE_0.2)

Reply: We added the abbreviation $KGE_{0.2}$ to the sentence.

Modification: L210

L210: Isn't the symbol "KGE" just good to indicate "no bias term was considered"? So, here could be KGE_0.5, similarly to KGE_0.2 in Eq.(1).

Reply: Yes, we could indeed simplify the abbreviation here. Following a suggestion from the other reviewer, we wrote the equation for O_{ETa} in full form. This way we can avoid introducing the label KGE.mod.

Modification: L220-221, equation 2

L230: How many grid cells in total? How do you define the increment of each axis?

Reply: We divided each axis into ten grid cells resulting in 100 cells in total. We added this information to be more precise about the percentage calculation.

Modification: L242-244

L252: Need to cite at the end of the sentence.

Reply: We added citations of Arciniega-Esparza et al. (2022), Dembélé et al. (2020a), and Mei et al. (2023) to support our statement.

Modification: L269-270

L265: I suggest to label this point too.

Reply: We updated the figure with a label for the Litchfield catchment.

Modification: Fig. 3

L265: I suggest to provide the flux map too as another example.

Reply: Adding the flux map for the Cow Bay catchment is a good idea. We updated the figure accordingly and added a couple of sentences to describe the flux map of the Cow Bay catchment.

Modification: L287-289, Fig. 3

L303: Is it just a random selection or is there some rules to follow when selecting the pairs?

Reply: Please see reply to comment 2.

Modification: L325-326

L311: Is there a more quantitative way to describe "separation" here? If yes, perhaps the authors can consider something like a pairwise CC plot that the X and Y axis are the seven parameters and the corresponding grids are the "separation metric". By doing so, all 21 pairs can be visualized and they can easily repeat this plot for the other catchments to make a 4X3 plot matrix.

Reply: Thanks for this very specific input. We visualized the quantitative results of the separation (see response to Comment 2) as suggested with a 4 x 3 plot matrix in Fig. 6 of the revised manuscript. We used Fig. 6 in addition to the existing 2-D scatterplot in Fig. 5 rather than replacing it, because Fig. 5 shows in a simple and convincing way the lack of overlap for some parameter distributions.

Modification: L323-328, Fig. 6

L326: In my opinion, the pagination of this manuscript is not very long. The authors may consider to make this 4X3 for all the ternary plots to occupy an A4 paper entirely. I don't think the last one of the current panel b is needed.

Reply: As suggested, we made the ternary plots (and the scatterplots) of the individual catchments bigger. We also adapted the figures in the supporting information accordingly. We didn't remove the last figure in panel b because it seemed relevant in the other reviewer's view.

Modification: Figs. 5 and 7, Figs. S11-12

L332: "Less constrained flux maps" and "increased hydrograph uncertainties" seem to be the same thing?

Reply: Since the hydrograph is a result of different combination of fluxes, we argue that increased hydrograph uncertainties are a result of less constrained flux maps. Although the two are inherently linked to each other, they have a slightly different meaning. We slightly modified the sentence to say that more explicitly: "… less constrained flux maps and *consequently* increased hydrograph uncertainties…".

Modification: L372

L334: "Improved overall model performance" does not necessary mean reduced uncertainties.

Reply: Thanks for pointing this out. We removed the "improved overall model performance" from the sentence to focus on improved realism.

Modification: L374

L372: This is a good point. But the performance associated with the 100 parameter sets were not shown in the manuscript. I suggest to add this piece of information.

Reply: Please see reply to comment 1.

Modification: L423-430, Fig. 8