Response to anonymous referee #1 on HESS egusphere-2025-3373 '*Using evapotranspiration signatures to assess evapotranspiration realism in rainfall-runoff models*'

Citation: https://doi.org/10.5194/egusphere-2025-3373-RC1

<table>
<tr><td>#</td><td>Remarks to author</td><td>Authors' responses</td></tr>
<tr><td colspan="3">General comment</td></tr>
<tr><td>1</td><td>This manuscript addresses an important topic in hydrological modeling by exploring joint calibration of rainfall–runoff models using both streamflow and flux tower evapotranspiration (AET) data. The study is relevant and has the potential to contribute to improving process representation and multi-variable calibration approaches. However, the paper would benefit from clearer methodological explanations and stronger conceptual consistency. The rationale for including the annual variability signature (CVannual) is unclear given that the adopted objective function explicitly ignores the annual timescale, and the description of calibration periods for discharge and AET is confusing and should more clearly distinguish between the fixed simulation window and site-specific data availability. Additionally, the use of the Nash–Sutcliffe Efficiency (NSE) to evaluate spatial patterns across catchments is not conceptually appropriate, as NSE is designed for temporal rather than spatial performance assessment. Some statements (e.g., regarding over- or underestimation of interannual variability) are internally inconsistent and should be revised for clarity. Overall, while the study is timely and methodologically relevant, it requires better alignment between the objectives, metrics, and interpretations to fully realize its potential contribution to the field. I hope the detailed comments provided in the attached pdf will help improve the</td><td>Thank you for your thorough and constructive review of our manuscript. We appreciate your positive assessment of the relevance and potential contribution of this study, as well as your detailed suggestions for improving its clarity, conceptual consistency, and scientific interpretation.

We have carefully considered your comments and provide our responses below. Based on these responses, we propose to revise the manuscript accordingly.

In particular, we have provided responses regarding:

- Clarification of the study objectives, metrics, and interpretation of the results and discussion.
- Clarification of the calibration periods for streamflow and AET.
- The rationale for using NSE to assess the agreement between observed and simulated AET signatures across catchments.
- Correction of inconsistencies identified in the manuscript.

Please note that the red texts within the quotation marks in authors' responses below indicates suggested revisions to the text in the revised manuscript.</td></tr>
</table>

	clarity, structure, and scientific rigor of the manuscript.	
Comments		
2	Introduction L56-62: A separate paragraph is necessary with clearer explanation of what is the aim of the study	Yes, we agree. Thank you for the suggestion. To make the aim of the study clearer, we suggest the below sentence to start last paragraph of the introduction. "The aim of this study is to examine the performance of calibrated rainfall-runoff models by evaluating the realism of simulated AET using the AET signatures. To achieve this, we are calibrating models in a way that is broadly similar to previous studies (including streamflow only calibration and streamflow jointly with AET, but never directly to the signatures), and then evaluating the realism of results through the lens of AET signatures…."
3	Section 2.3: In this section, can you explain why these signatures are selected? and what do they represent exactly?	Yes, we agree some explanation would be beneficial. We suggest adding the following text at the beginning of Section 2.3. "The AET signatures defined by Gardiya Weligamage et al. (2025) provide a set that is reasonably representative of the variety of characteristics inherent to AET dynamics across different temporal scales, which are ideal for assessing the realism of AET representation in conceptual rainfall-runoff models. Collectively, they characterise long term average and median, inter- and intra-annual variability, seasonal behaviour, AET responses to rainfall and water limitation. Table 2 lists the AET signatures defined by Gardiya Weligamage et al. (2025). Most of these signatures are self-explanatory, except for two…."

4	Section 2.5 L1401-142: can you give examples of how these selected model structures provides different model structure complexities and et representations? Also it would be very beneficial to give a figure of these model structures (how they looks).	Yes, we agree that providing information on the selected model structures and AET representation is beneficial. We would like to suggest including a summary table describing model structures and AET representation of the selected models, as adding schematic diagrams of model structures would complicate the manuscript content. We will also guide readers to appendix of the MARRMoT toolbox manual (used in this study for the modelling), where schematic visualization of the mode structures are provided.
5	Section 2.5.2 L183-184: Please clarify the last sentence. It is currently unclear whether different calibration periods were used for discharge and AET or whether the simulation period also varied by site.	Yes, we agree Thank you for pointing out the sentence is unclear. We suggest an update text like below. "Here, similar to Method 1 in section 2.5.1, the simulation period was selected from 1980-2022, and model calibration using discharge was conducted from 1982-2022. However, model calibration using flux tower AET was conducted over time periods that varied from site to site, depending on the availability of flux tower data (Table 1). Where flux tower AET data extended beyond 30 April 2022 (the final date of streamflow data), only data up to this date were used to maintain a common simulation period between the two methods."
6	Section 2.5.2 L189: why a threshold of 0.6 was selected as the optimised objective function? Can you justify your choice? Why not 0.5 or 0.7?	Yes, we agree. Thank you for this comment as we agree that the rationale for selecting the threshold should be clarified. Therefore, we suggest the following updated text.

		"Regardless of the calibration method, the simulated AET was only evaluated in this study if the model met pre-set performance criteria. The idea is to focus on instances where performance (judged by traditional metrics) is broadly acceptable and use the signatures to gain insight into failings that are not apparent from those traditional metrics. For this purpose, a threshold of 0.6 was adopted for the optimised objective function in both calibration methods. Such a threshold was inherently subjective; a higher threshold would have been desirable but would have left the study with far fewer models to analyse. The value of 0.6 screened out the lowest most while still allowing for a reasonable sample of models to analyse. …."
7	Section 2.6 L200: Before directing readers to supplementary material, it would be good to shortly explain what is the result they found - whether if the results change when focusing on a subset of data or not.	Yes, we agree. We suggest the following updated text. "For this purpose, a subset of study sites was further selected based on the closer proximity of flux towers (either very close to or within the catchments) and similarity in dominant land cover for a similar comparison of AET signatures between models and flux towers but to get insights into whether the results change when focusing on a more representative subset of the data. This analysis confirmed that the results were broadly consistent with those obtained using all 14 sites, with no substantial changes in the overall findings. The results are available in Section S3 in the Supplementary material."
8	Section 3.1 L216: this is relatively wide range of increase in scores due to addition of AET into the calibration. Can we explain these differences in	Yes, we agree. We will discuss this under Discussion. As shown by the results,

	the increase by the differences in model structures used? Are these differences expected based on the model structures used?	the more structurally complex Sacramento model showed the smallest increase in AET performance (increase in median value of approximately 0.06), whereas SIMHYD showed the largest increase (increase in median value of 0.41). This suggests that the magnitude of improvement is not simply related to overall model complexity, but may instead depend on how soil moisture and evapotranspiration processes are represented within each model. For example, the relatively simple soil moisture accounting in SIMHYD may allow AET observations to more effectively constrain model parameters, whereas the more complex soil moisture representation and larger number of interacting parameters in Sacramento may increase parameter equifinality, potentially limiting the additional information gained from incorporating AET observations into calibration. We will further discuss these model-structure-related differences in the revised manuscript.
9	Section 3.1 L220: Can you justify this claim by giving some references in previous section where you stated that 0.6 is the chosen threshold for the performance evaluation?	As responded in comment #7, the threshold of 0.6 was adopted for the optimised objective function in both calibration methods to represent reasonably acceptable value given the reasons states above. Since it is a subjective choice, we will clarify the scores we considered as "good" is above 0.6.
10	Section 3.2.1 L238: The statement about interannual variability (CVannual) appears internally inconsistent. The text first mentions "overestimation" relative to flux tower estimates but then states that variability in reality is higher than in simulations, which would imply underestimation. Please clarify this.	Thank you so much for carefully reviewing the manuscript. We acknowledge that was an error in the text. We will change the sentence as below. "The interannual variability (expressed as the coefficient of

		variation, $CVannual$) shows overestimation compared to flux tower estimates (Figure 4); in other words, the variability in annual AET is higher in the simulations than in reality."
11	Section 3.2.1 L242: It is stated that including daily and monthly AET information in the calibration does not improve annual AET dynamics because the adopted objective function ignores the annual timescale. If this is indeed the reason, it is unclear why the annual-scale signature (CVannual) was included as a performance signature, especially since the authors also note that such objective functions typically ignore annual dynamics. Please clarify the rationale for selecting CVannual as a diagnostic metric in this context	Thank you for this important comment. We agree that the rationale for including CVannual should be explained more clearly. The purpose of including CVannual was not because it was expected to improve directly through calibration, but because it represents an important aspect of AET realism that is independent of the adopted objective function. Evaluating signatures that are not explicitly targeted during calibration allows us to assess whether improvements achieved at the daily and monthly timescales translate to the annual scale. Although it may be argued that annual metrics should be included in the objective function if they are important to the context, this belies common practice wherein modellers default to common objective functions without thinking about the match with study aims. By reporting these results, we hope to encourage reflection such that objective functions are more carefully selected in future. The finding that CVannual shows no improvement is therefore itself an important result, demonstrating that improvements in shorter-timescale AET dynamics do not necessarily propagate to longer timescales. We will revise the manuscript to clarify this motivation.
12	Section 3.2.2 L250-251: Why SIMHYD and GR4J tend to show improvements, not other models? Any thought on this?	Thank you for this comment. We think that some improvements observed in the timing of seasonal peak of the SIMHYD and GR4J suggest that incorporating AET observations was more effective in

		correcting the seasonal timing of evapotranspiration in these models. This may be related to differences in the representation of soil moisture dynamics and evapotranspiration processes, which influence the partitioning of water availability and evaporative demand throughout the year. We will discuss these possible structural explanations in the revised manuscript.
13	I would suggest you to give Figure 3 and 4 at the end of paragraph in the section 3.2.1 and Figure 5 and 6 at the end of paragraph in the section 3.2.2. It would be easier for readers to follow.	Yes, we agree. We will change the placement of the figures as suggested.
14	Figure 10: I would give this figure at the end of paragraph at Section 3.2.4.	Yes, we agree. We will change the placement of the figure as suggested.
15	Section 3.3: L 307-311: NSE is calculated between observed and simulated signature values across 14 catchments. However, NSE is primarily designed to evaluate temporal performance at individual sites rather than spatial agreement across catchments. For spatial comparison of signatures, correlation coefficients or normalized RMSE-based metrics would be more conceptually appropriate. Please clarify the rationale for using NSE in this context or consider adopting a spatial performance measure.	We agree that NSE is typically used to evaluate temporal performance at individual sites, but it can also be framed as a normalised RMSE metric since it can be expressed as 1 minus the squared ratio RMSE/stdev. Correlation coefficients are not suitable here since they do not account for bias, and raw RMSE values are difficult to intercompare among variables with different magnitudes. Hence, we feel the NSE is an appropriate measure.