Responses to Anonymous Referee #1

This study investigates the capacity of the VIC model to integrate streamflow and evaporation data in a large-sample application, based on simulations for 189 headwater catchments in Spain. Utilizing multiple datasets to improve model performance is an important aspect of hydrological modeling research, making this paper potential for publication in HESS. However, some important issues are not discussed adequately, and there is room for improvement. Consequently, I recommend a major revision before publication.

We thank the reviewer for his/her constructive feedback and we are convinced it will contribute to improve the manuscript. We have indicated in our responses those references that were not included in the initial version of the manuscript.

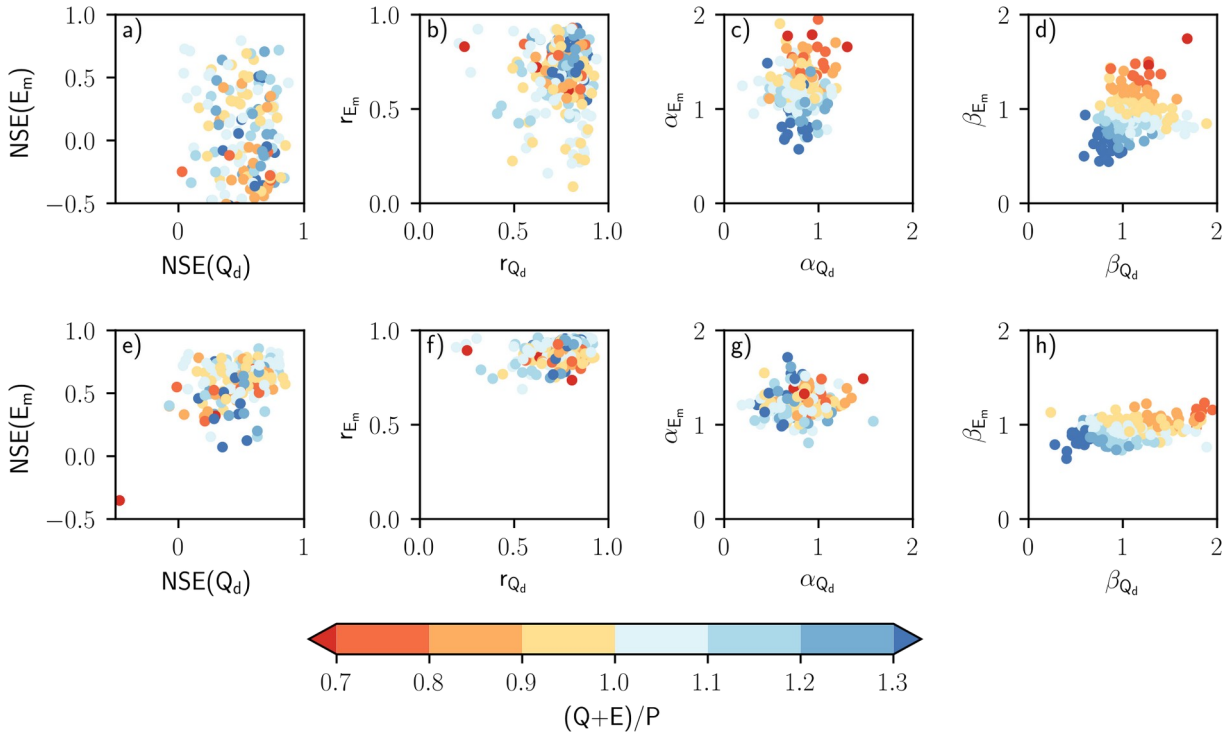
My major suggestions are as follows:

Given the general imbalances in Q, E, and P, I question whether it is reasonable to use the evaporation dataset to evaluate the model directly. This brings another question: what is the most important signature provided by the evaporation data? Is it the evaporation amount or the temporal variation of evaporation? If the key information provided by the evaporation data is the amount, we can expect that the E simulation obtained by Q-only calibration will be good in the catchments with Q+E/P close to 1, and the NSE(Q) obtained by Q-E calibration will be lower than that from Q-only calibration (perhaps the authors can test whether the results indeed show this characteristic). Otherwise, if the key information provided by the evaporation data is the variation, conducting a bias correction on E data based on water balance before calibration makes more sense.

We thank the reviewer for pointing out this important issue regarding the imbalances in Q, E and P. In our understanding, the key information provided by the evaporation dataset is the amount of evaporation. This does not mean that the temporal dynamics are not important, but they are evaluated at the monthly time scale and thus are less relevant than in the case of streamflow, which is evaluated at daily scale. The amount of evaporation together with streamflow and precipitation allow for identifying gaining and losing catchments as they constitute an indirect measure of the intercatchment groundwater flow (Liu et al., 2020). Gaining and losing catchments are therefore characterized by an unclosed water balance, which can potentially lead to an unrealistic representation of the partitioning of precipitation into streamflow and evaporation in case of significant imbalances when using a (closed) water balance hydrologic model.

This issue was addressed in our original submission by performing two calibration experiments, namely Q-only and Q-E calibration, in order to assess the relative gain/loss in model performance in terms of NSE(Q) and NSE(E). Results showed that NSE(Q) remained similar for both calibrations, indicating that the imbalances in Q, E and P did not deteriorate model performance for streamflow in a substantial manner. However, we agree that there is room for improvement and that this issue is worthy of further examination. As suggested by the reviewer, we have evaluated the performance for Q and E during both calibration experiments considering the (Q+E)/P ratio as a signature of how far/close is the

water balance from being closed from a data perspective. We have performed the analysis for NSE as well as for its decomposition into r, α, and β components for the 189 study catchments (please, see also our response in relation to the decomposition of NSE into r, α, and β). The results are depicted in the following figure:



Panels a-d) correspond to the Q-only calibration and panel e-h) correspond to the Q-E calibration, and the different values have been calculated for the complete study period. As rightly pointed out by the reviewer, the NSE(E) estimate obtained with the Q-only calibration is better for the catchments with (Q+E)/P close to 1 (panel a), and the NSE(Q) produced by the Q-E calibration is lower than that from the Q-only calibration (panel e). As suggested in Yeste et al. (2023), the β component (i.e., the bias component) is of capital importance from a water balance perspective as it is sensitive to the imbalances of the Q, E and P data when Q and E are integrated into model calibration. As shown in panels d) and h), β is closer to 1 for both Q and E for catchments with (Q+E)/P close to 1, with a wider distribution of β_Q for the Q-E calibration due to the imbalances of the Q, E and P data. These imbalances, however, do not have a marked effect for the dynamics (i.e., r) and the variability (i.e., α), as shown in panels b) and f), and c) and g), respectively. Hence, although NSE(Q) values are slightly lower for the Q-E calibration, they remain similar for both calibration experiments, as indicated before.

This figure will be introduced in Section 4.2 and will be discussed in Section 5.2 in the revised version of the manuscript.

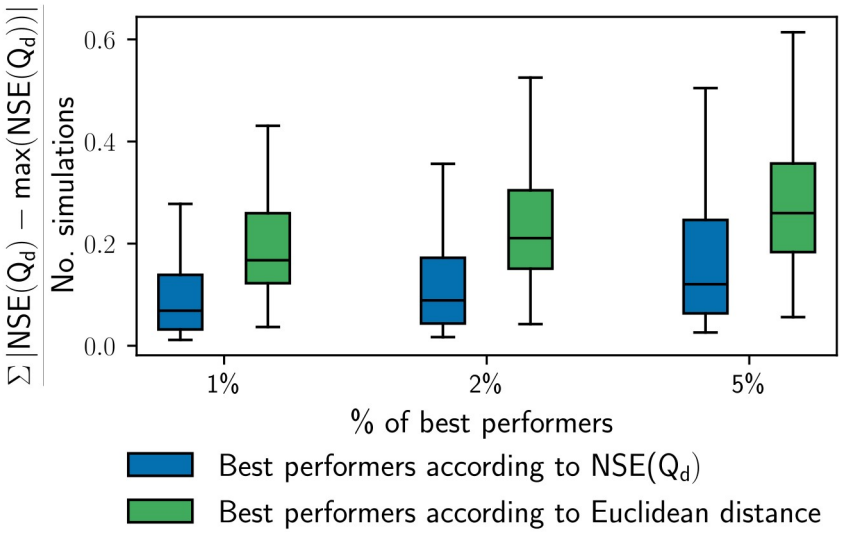
Liu, Y., Wagener, T., Beck, H. E., & Hartmann, A. (2020). What is the hydrologically effective area of a catchment? Environmental Research Letters, 15(10). https://doi.org/10.1088/1748-9326/aba7e5

An important role of adopting multiple datasets is to reduce equifinality, i.e., to reject some parameters that perform well in the simulation of only one objective. However, the authors didn't address this issue, stopping at presenting good simulations for streamflow and evaporation. I encourage the authors to discuss the value of the evaporation dataset in reducing equifinality. A potential way to address this issue is to analyze the sensitivity of the integrated objective function to the parameters and compare it with the sensitivity of NSE(Q).

This is ideally done following a Pareto optimization approach for both NSE(Q) and NSE(E) as in Yeste et al. (2023). Pareto optimization is known for reducing equafinality as the model is optimized for two or more objective functions simultaneously, resulting in a lower number of behavioural parameter sets (Efstratiadis and Koutsoyiannis, 2010). Although we have not implemented a full Pareto optimization, it is still possible to address this issue only considering the two calibration experiments in this work as they represent two important solutions belonging to the Pareto front: the corner corresponding to the maximum NSE(Q) performance and the compromise solution for NSE(Q) and NSE(E) considering a weighted Euclidean distance with equal weights for both.

We value the suggestion from the reviewer to compare the sensitivity indices of NSE(Q) and the integrated objective function. However, the sensitivities indices are based on the Monte Carlo experiment comprising 10000 Latin Hybercube samples, and therefore they are representative of the complete parameter and objective space. As equafinality is concerned with behavioural parameter combinations, we propose that such comparison should be carried out only for the region where the best candidate solutions are located.

Following this reasoning, we have calculated for every catchment the mean absolute deviation of NSE(Q) from the maximum NSE(Q) considering the 1%, 2% and 5% best performing simulations from the Monte Carlo experiment and according to two criteria: 1) NSE(Q) itself and 2) the Euclidean distance for NSE(Q) and NSE(E). Results are shown in the following figure:



The boxplot above shows two effects: 1) as the percentage of best performers considered increases, the deviations from the maximum NSE(Q) become higher; 2) the deviations are

more pronounced when the best performing criterion is based on the Euclidean distance for NSE(Q) and NSE(E). The first effect is a straightforward consequence of considering an increasing number of simulations to calculate the mean absolute deviation from the maximum NSE(Q). The second effect is an indicator of less equifinality as there are fewer parameter combinations yielding a performance close to the maximum NSE(Q), which highlights the value of using multiple datasets in reducing equifinality.

This figure will be introduced in Section 4.1 and will be discussed in Section 5.1 in the revised version of the manuscript.

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal, 55(1), 58–78. https://doi.org/10.1080/02626660903526292

Other minor and moderate issues:

- Data

There are negative records at some stations, which can be attributed to the reservoir. So a question is whether the influence of the reservoir on streamflow is significant and whether the reservoir is simulated in the model. To my knowledge, some rivers significantly influenced by reservoirs have an extremely even interannual streamflow distribution, which is impossible to reproduce if the reservoir is not considered in the model.

The streamflow dataset is not affected by river regulation as the study catchments selected in this work are located in headwater areas and there are no dams within them. Some of the catchments were delineated considering the reservoirs as outlets, and these are the stations that the reviewer is referring to. As described in Section 2.1, the streamflow time series for reservoirs were indirectly estimated as a water balance between water storages and releases. The regulated component corresponds to the water release, which is subtracted from the water storage to calculate the input flow to the reservoir. The latter is then equated with the streamflow of the catchment, and therefore there is no influence of the reservoir on streamflow as the input flow to the reservoir is not regulated.

- Methods

The model performance was evaluated using NSE and its decomposition into γ, α, and β. However, to my understanding, γ, α, and β are the decomposition of KGE rather than NSE. NSE actually only quantifies the bias characteristic. I suggest the authors modify this expression. Additionally, please add the equations for these three metrics.

Both KGE and NSE can be decomposed into r (correlation coefficient), α (variability term), and β (bias term), and therefore both constitute integrative metrics of model performance. Thus, NSE not only quantifies the bias characteristic (i.e., β), but also the dynamics (i.e., r) and the variability (i.e., α). The expression for NSE as a function of r, α, and β is given by (e.g., Knoben et al., 2019):

$$NSE = 2\alpha r - \alpha^2 - \frac{(\beta - 1)^2}{CV_{obs}^2}$$

where CV_{obs} is the coefficient of variation of the observations. The equation of NSE as a function of r, α, and β and the equations for α, and β will be indicated in the revised version of the manuscript. The correlation coefficient does not need to be defined given its widespread use.

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences, 23(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

I am a little confused about the SST test. If I understand correctly, this seems to be the common practice in model calibration, i.e., to divide calibration and evaluation periods, with a warm-up period in each. Please correct me if I am wrong, but if this is indeed the common practice, I suggest not referring to it using a special term.

The pioneering work of Klemeš (1986) defines two splitting strategies: the Split-Sample Test (SST) and the Differential Split-Sample Test (DSST). Both approaches have been extensively used in hydrology, and the terms SST and DSST are specifically referred to in a large number of hydrologic studies (e.g., Fowler et al., 2018, 2021; Gharari et al., 2013; Melsen et al., 2019; Rakovec et al., 2019) as it is a direct way to indicate which calibration approach has been applied. Therefore, we consider that it is important to use the term SST as it can help the hydrologic community to clearly identify the practice followed in our study. In alignment with all the previous studies in which this terminology is specifically used, we have decided to keep our reference to the SST test.

Fowler, K., Coxon, G., Freer, J., Peel, M., Wagener, T., Western, A., Woods, R., & Zhang, L. (2018). Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement. Water Resources Research, 54(12), 9812–9832. https://doi.org/10.1029/2018WR023989

Fowler, K. J. A., Coxon, G., Freer, J. E., Knoben, W. J. M., Peel, M. C., Wagener, T., Western, A. W., Woods, R. A., & Zhang, L. (2021). Towards more realistic runoff projections by removing limits on simulated soil moisture deficit. Journal of Hydrology, 600(December 2020), 126505. https://doi.org/10.1016/j.jhydrol.2021.126505

Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2013). An approach to identify time consistent model parameters: sub-period calibration. Hydrology and Earth System Sciences, 17(1), 149–161. https://doi.org/10.5194/hess-17-149-2013

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Zappa, M., Mizukami, N., Mendoza, P. A., Clark, M. P., & Uijlenhoet, R. (2019). Subjective modeling decisions can significantly impact the simulation of flood and drought events. Journal of Hydrology, 568(September 2017), 1093–1104. https://doi.org/10.1016/j.jhydrol.2018.11.046

- Sensitivity Analysis

The authors calculate the sensitivity of model performance to parameters and analyze the correlation between sensitivity and physiographic hydroclimatic characteristics. I think it would be interesting to discuss the mechanism behind this correlation, e.g., why NSE(Q) is more sensitive to rout1 in catchments with larger precipitation. This can also provide guidance on selecting sensitivity parameters in regions with different conditions. I

encourage the authors to delve deeper into this analysis or discussion. Currently, this is only discussed by comparing with other studies, without addressing the underlying reasons.

We thank the reviewer for his/her suggestion. We think that investigating the mechanisms behind the correlations between the parameter sensitivities and the physiographic and hydroclimatic characteristics will positively contribute to improve the manuscript. The following discussion revolves around two axes: 1) NSE(Q) sensitivities, 2) NSE(E) sensitivities.

- NSE(Q) sensitivities: as stated in Section 4.1, the highest NSE(Q) sensitivities correspond to the soil parameters (b_i, D_S, W_S, D_m and d_2) and the routing parameters ($rout_1$ and $rout_2$) (Fig. 4). The soil parameters presented an opposite pattern (i.e., negative correlations) to mean annual precipitation, aridity index, NDVI and to a lesser extent slope (Fig. 6). These parameters control the runoff generation process in VIC, and the negative correlations indicate that they are more important for catchments characterized by a more arid climate. As the precipitation volume to be transformed into runoff is lower for such catchments, the role of the five soil parameters becomes critical in modulating the runoff generation, whereas their effect is less relevant for catchments belonging to a more humid climate given the higher water availability. The generated runoff volume is subsequently routed to the catchment outlet according to a gamma-based unit hydrograph in a post-processing phase. The two routing parameters control the delay between runoff generation and catchment discharge (i.e., streamflow) and exhibited a matching pattern (i.e., positive correlations) to the previous four attributes (Fig. 6), suggesting that both parameters are important for the humid catchments as a consequence of the higher runoff volumes to be routed.

- NSE(E) sensitivities: NSE(E) was greatly influenced by the vegetation parameters, in particular by $rmin_f$ and LAI_f (Fig. 5). Among the soil parameters, d_2 was revealed as the most important soil parameter to NSE(E), and as expected, the routing parameters showed a null effect. The high NSE(E) sensitivities for $rmin_f$ and LAI_f reflected negative correlations to mean annual precipitation, aridity index and NDVI (Fig. 6), denoting a greater impact for arid catchments that is likely associated to the limiting effect on the evaporative processes entailed by a lower vegetation density. As for the soil parameters, the high NSE(E) sensitivities for d_2 could be related to the water uptake by vegetation in the root zone as it is directly affected by the thickness of the VIC soil layers. The positive correlations associated to the five soil parameters with respect to the previous characteristics are likely connected to the implementation of the closed water balance equation VIC and manifest an opposite behaviour to that observed NSE(Q). This effect was also appreciated in Yeste et al. (2020) for the sensitivities of the VIC soil parameters.

The previous discussion will be integrated in Section 5.1 in the revised version of the manuscript.

The Y-axis of Figure 7 is incorrect.

We thank the reviewer for pointing out this mistake. It will be modified in the revised version of the manuscript.