Responses to Ilhan Özgen-Xian

Summary. This article investigates how using evapotranspiration in addition to streamflow data during the calibration process affects the model performance across 189 Spanish catchments. The Variable Infiltration Capacity (VIC) model, a semi-distributed hydrological model, is used for modelling. For the calibration of evapotranspiration fluxes, leaf area index and minimal stomatal conductivity were found to be the most sensitive parameters. Apart from the vegetation parameterisation, soil parameters were found to have a large effect on model results. The VIC model is run for 20 years (with a spin-up period of 10 years) using different targets for the calibration. Results are analysed and discussed with a focus on model sensitivity to calibration target.

Assessment. The subject of this article is interesting. Indeed, large sample and large scale hydrological modelling is becoming more feasible, due to improvements in computing technology. In this sense, the topic is timely and of interest. The manuscript is well written and easy to follow. I think the modelling work is substantial and the work is suitable to be published in HESS. I have some questions that I would like the authors to address. These are listed below. I recommend minor revision before publication.

We thank the reviewer for his positive feedback and interesting questions as they 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.

Comment on large sample modelling in this work.

The use of large samples in hydrology is very interesting. Especially when combined with a comparative analysis, the large sample size can help us to discover relations among processes and generate hydrological insights. While this work made use of data from a large number of catchments, the discussion of the results were focused on quite technical issues of model structure and calibration. I do not think this is bad, in fact, these are important topics. But I wonder whether the conclusion in the abstract, i.e. "This investigation will help gain a better understanding of the hydrology of the Spanish catchments and will help prepare the ground for a fully gridded implementation of the VIC model in Spain." still holds. Perhaps in the conclusions, the authors could provide some insights of the hydrology of Spanish catchments they gained in this study. Or perhaps this manuscript focuses on model sensitivity, in this case, this should be better reflected in the abstract.

We thank the reviewer for his positive feedback regarding the use of large samples in hydrology. Indeed, the importance of large-sample hydrology lies on the analysis of multiple catchments to draw general conclusions about the hydrologic functioning of the domain under investigation. We expect that the results from this work will be the basis for a future implementation of VIC for the Spanish catchments and will contribute to produce seamless distributed parameters maps and Spanish-wide simulations based on a fully gridded implementation. However, this is a future development and not a current result from this work. Therefore, we apologize for concluding in the abstract in such a way. As the manuscript focuses on model sensitivity and model performance during calibration, this will be better reflected in the abstract in the revised version of the manuscript.

Questions.

1. Here is my understanding of the modelling work, please correct me if I am wrong: The VIC model is a semi-distributed hydrological model in the sense that no horizontal fluxes are computed between individual grid cells. VIC is set up for entire Spain, thus, all catchments included in this studies data set are represented in the model. Model calibration is done by adjusting model parameters in each grid cell individually.

The description of the VIC model is correct. However, VIC was not implemented for entire Spain, but rather for the individual catchments with a spatial resolution of 5 km for the grid cells overlapping the catchment area. As for the calibration, the calibrated parameters were considered as spatially constant for all the cells affecting the catchment. The rest of the soil and vegetation parameters not modified during calibration were derived for each cell individually. This approach was also followed in previous implementations of VIC in Yeste et al. (2020, 2022, 2023).

2. How were soil hydraulic properties aggregated from 1 km to 5 km?

As described in Section 3.1, soil parameters were regridded to the model resolution (i.e., from 1 km to 5 km) using a first-order conservative remapping, which corresponds to an area-weighted average of the soil properties.

3. Streamflow and evapotranspiration processes have distinct time scales. Are model results of the same variable at different temporal resolution, for example, daily vs. subdaily stream flow, sensitive to different model parameters?

Parameter sensitivities were exclusively analyzed with respect to the NSE of daily streamflow and the NSE of monthly evaporation as these were the objective functions that were used during calibration. In this sense, the Regional Sensitivity Analysis was performed as a previous and necessary step to the calibration that allowed for identifying the most important parameter with respect to both metrics, which were later selected as the calibration parameters. Although the effect of the temporal resolution on parameter sensitivities has not been explored in this work, it is important to mention that parameter sensitivities may indeed vary depending on the temporal resolution of the variable analyzed. For instance, Melsen and Guse (2019) showed in a large-sample application of VIC over the CONUS domain that the effect of LAI_f and $rmin_f$ is negligible for the NSE of daily streamflow but becomes more important at annual scale.

4. Does calibration with only streamflow, only evapotranspiration, and both of them combined result in significantly different model parameterisations?

Please note that only two calibration experiments were performed in this study: a calibration using streamflow data (Q-only calibration) and a calibration combining streamflow and evaporation data (Q-E calibration). No calibration was performed for evaporation exclusively. This is an interesting question and is closely connected to another question posed by Anonymous Referee #1. One of the main advantages of using multiple datasets for model calibration is to reduce equifinality. This implies a decrease in the number of behavioural parameter combinations as two or more datasets are used to constrain the model. As part of our response, we created a new figure that will be introduced in Section 4.1 and will be discussed in Section 5.1 in the revised version of the manuscript. Please refer to our response to Anonymous Referee #1 regarding the benefit of including evaporation data to reduce equifinality.

5. Using the model results in this study, can the relations between the Nash-Sutcliffe efficiencies and model parameters reported in Fig. 6 be interpreted from a hydrological point of view?

Yes, the correlations reported in Fig. 6 can be interpreted from a hydrologic perspective and it will be done in the revised version of the manuscript. This same question was posed by Anonymous Referee #1 and was extensively discussed in our answer, so please refer to our response to Anonymous Referee #1 for further details. We are convinced that the discussion on the mechanisms behind the correlations will greatly contribute to improve the manuscript and will be integrated in Section 5.1 in the revised version of the manuscript.

6. On page 11, it is reported that a simultaneous calibration with both streamflow and evapotranspiration results in a degradation of model performance. Hydrological models that have been calibrated against more than one type of data often display a greater generalisation capability to changing climate conditions. Can this be seen for the VIC model in the simulated time frame in this study? Is this what is discussed in the last paragraph of Sec. 5.2?

Please note that the deterioration of model performance that is reported on page 11 only refers to streamflow, as the performance for evaporation highly increases when the model is simultaneously calibrated with both streamflow and evaporation data (Fig. 8, 9). The generalization capabilities of the VIC model to changing climate conditions have been assessed by comparing the model performance for streamflow and evaporation for the calibration and the evaluation periods during the Split-Sample Test (Fig. 9). However, the loss in model performance during the evaluation period when compared to the calibration period was similar for both the Q-only and the Q-E calibration experiments, and therefore this effect was not visible based on the results of this study. This could be better explored by implementing a Differential Split-Sample Test (Klemeš, 1986) after selecting two contrasting periods as in Fowler et al. (2018). We acknowledge this is an interesting approach and a potential future development to further test the predictability of the VIC model. The last paragraph of Sec. 5.2 refers to the cross-validation test performed using different observational datasets of precipitation and temperature for the study period as a way to assess the generalizability of the calibrated parameters.

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

7. Sec. 3.3: What climate forcing was used to spin-up the simulation? From the corresponding 10 years preceding the simulation period?

We used daily precipitation and temperature data for the 10 years preceding the simulation period (i.e., spin-up period).