Dear Referee #1,

We would like to thank you for your review of our paper and your fruitful comments on it. Please find below our replies (plain text) to your comments (in italic) with intended changes to the manuscript.

This manuscript presents different calibrations and evaluations based on different climatic conditions spatially and temporally intended to explore how the model performance are sensitive to drought condition and the real potential causes data, and test that a drought included in calibration period would improve model transferability or not. The authors designed two calibration experiments and three evaluations under three different wetness conditions. They mainly found that a drop in performance of Q indeed happened in their study area based on Continuum model, and was related to the representation of ET anomalies rather than TWSA, and including a moderate drought in the calibration did not lead to an improvement in Q and ET simulation during a severe drought. The research makes a contribution to understanding the application of LSASAF product, and model performance under different climatic conditions.

However, I have some major concerns that need to be addressed prior to reconsideration:

1) Based on the results, I don't think authors can say that "the drop in Q modelling performances during the severe 2022 drought event can be related to the mis-representation of ET anomalies, among other factors". I just observed that a drop happened in ET performance from moderate droughts to severe droughts which was similar with that for Q performance. I didn't see any other evidence to prove that drop in Q performance can be related to ET simulation. Please design more experiments to give audience more evidence.

The drop in the simulation of streamflow (Q) during the severe drought was mainly driven by an overestimation of Q in evaluation sub-catchments. To clarify this point, we will adjust Figure 4 by grouping the boxplots in panel d into calibration and evaluation sub-catchments, and we will add in the supplement similar figures for the components of the Kling Gupta Efficiency (correlation, bias, and variability). We will add further analyses to explore the potential causes for this. Specifically, we will investigate (i) systematic biases in observed data that may have occurred during the severe drought, because of increased uncertainty in P data or enhanced human disturbance on Q data for instance, (ii) the overestimation of the simulated contribution of Terrestrial Water Storage (TWS) to Q generation, and (iii) the underestimation of simulated evapotranspiration (ET). The uncertainty in observed data we used to force and evaluate the model has not increased systematically during the severe drought (69±234 mm in 2012 and 51±202 mm in 2022 as mean ± 1 standard deviation across the study sub-catchments). Furthermore, at the outlet section the model slightly overestimated TWS during the severe event (simulated TWS = -65 mm vs observed TWS = -92 mm in September 2021, and simulated TWS = -100 mm vs observed TWS = -158 mm in August 2022, Figure 5), and thus it did not overestimate its contribution to Q. Therefore, we conclude that the underestimation of ET was the main cause for the drop in Q performances, especially for human disturbed areas, as emerged also from the analysis on spatial patterns. Previous literature on the topic supports our finding. Avanzi et al. (2020) for instance identified the misrepresentation of ET elasticity to climate variability as the culprit for the drop in Q simulation during the 2012-2016 Californian drought for a semi-distributed hydrological model. We will revise Sections 3.2, 3.3, and 4.1 extensively to show and discuss these points.

2) I don't think the authors really solved the research question: would a drought in calibration period improve model transferability or not? Authors just have two calibration experiments(one with normal period and another one with moderate drought) and then compared the evaluated results in severe drought. The results in this study were very different from that in Yang et al. (2021). But this may result from the different model which Yang used and this reference can not prove your result is correct. I suggested that author can design more experiments including different type of droughts based on different models and compare their results so that make your results more reliable.

Yang et al. (2021) tested different calibration strategies for the simulation of the 2018-2019 German drought with an ecohydrological model in an experimental catchment and they reported an improvement in model performances by including the drought in the calibration period, compared to those from a wet calibration period. However, Avanzi et al. (2020) revealed that a semi-distributed hydrological model calibrated also during a drought had a drop in model performance when evaluating it during the 2012-2016 Californian drought. Here we calibrated a distributed hydrological model during a moderate drought (the 2017 event over the Po river basin), and then evaluated it during an independent and more severe drought (the 2022 event), without substantial improvements in model performances compared to those from an alternative calibration period. We agree that our conclusions differ from those in Yang et al. (2021) and, in our discussion, we indeed intended to refer to Yang et al. (2021) as a study contrasting our conclusions, rather than supporting them. This may be due to a number of differences between the two studies: first of all the experimental design (use of an independent drought as period for model evaluation), as well as differences in models, study areas, and calibration procedures used. We will clarify this by expanding lines 292-295 in Section 4.1. We agree that comparing different models in their transferability to severe droughts when calibrated during moderate droughts could provide interesting insights on the topic. While this is beyond the scope of the current paper, we will add it as a further possible way forward for future research in Section 4.2.

3) how do you define wet, normal, moderate drought, and severe drought? I didn't see detailed clarification or an indicator in this manuscript.

We characterized the different wetness conditions over the study period in terms of annual P standardized anomalies, as reported in Section 2.4.1 and Figure 2. We identified the wet/dry periods as periods with positive/negative anomalies for most of the study sub-catchments. Further, we referred to dry years as droughts, and we defined them moderate and severe in terms of decreasing annual P standardized anomalies. We did not set any specific threshold on annual P standardized anomalies to define drought years and characterize their severity; however, our drought characterization agrees with previous literature and drought reports (Masante et al., 2017; Marchina et al., 2019; Toreti et al., 2022a, b). We will clarify this in Section 2.4.1.

4) could you please add the evaluation performance results in supplementary?

Yes, we will add the evaluation scores for Q over all the evaluation periods for each calibration experiment and sub-catchment in a table in the supplement material.

5) please make your paragraph format consistency.

We will make our paragraph format consistent throughout the manuscript.

6) what does the grey shade represent in Figure 5? Please add that in the text blow the figure.

The grey shade in Figure 5 represents the analyzed drought years and we will specify it in the caption.

7) what does the river basin really look like? When I see the river basin in figure1 and 4, the river basin looks well, however, the river basin in figure6-8 looks like that it was stretched vertically. Please make the river basin consistency in your figures. And please organize your figure6-8 better.

We will modify Figures 6-8 accordingly and rearrange the subplots in them.

8) Line 203, what is "a climatology"? please clarify it in details.

With "climatology", we meant the mean ± one standard deviation over the study period (see caption of Figure 3). We will specify it also in the text in Section 3.1.

References:

Avanzi, F., Rungee, J., Maurer, T., Bales, R., Ma, Q., Glaser, S., and Conklin, M.: Climate elasticity of evapotranspiration shifts the water 370 balance of Mediterranean climates during multi-year droughts, Hydrology and Earth System Sciences, 24, 4317–4337, 2020

Marchina, C., Natali, C., and Bianchini, G.: The Po River water isotopes during the drought condition of the year 2017, Water, 11, 150, 2019

Masante, D., Vogt, J., McCormick, N., Cammalleri, C., Magni, D., and de Jager, A.: Severe drought in Italy – July 2017,https://doi.org/https://edo.jrc.ec.europa.eu/documents/news/EDODroughtNews201707_Italy.pdf, 2017

Toreti, A., Bavera, D., Acosta Navarro, J., Cammalleri, C., de Jager, A., Di Ciollo, C., Hrast Essenfelder, A., Maetens, W., Magni, D., Masante, D., Mazzeschi, M., Niemeyer, S., and Spinoni, J.: Drought in Europe August 2022, https://doi.org/doi:10.2760/264241, 2022a.

Toreti, A., Bavera, D., Avanzi, F., Cammalleri, C., De Felice, M., de Jager, A., Di Ciollo, C., Gardella, M., Gabellani, S., Leoni, P., Maetens, W., Magni, D., G., M., Masante, D., Mazzeschi, M., McCormick, N., Naumann, G., Niemeyer, S., Rossi, L., Seguini, L., Spinoni, J., and van den Berg, M.: Drought in Europe April 2022, https://doi.org/doi:10.2760/40384, 2022b.

Yang, X., Tetzlaff, D., Soulsby, C., Smith, A., and Borchardt, D.: Catchment Functioning Under Prolonged Drought Stress: Tracer-Aided Ecohydrological Modeling in an Intensively Managed Agricultural Catchment, Water Resources Research, 57, e2020WR029 094, 2021.