## Response by the Authors to Reviewer 2 for EGUSPHERE-2025-2478

We would like to thank the reviewer for the valuable recommendations to improve this manuscript and its title. The original reviewer comments are presented in black font, and our responses are in blue font.

In their manuscript, Sofokleous et al. aim to investigate the impacts of climate change on streamflow in Cyprus. To this end, the authors test various metrics for model calibration. While their general scientific aim is valid and fits well with HESS, the article's current scientific quality requires major revisions. A revision is necessary to address the article's framing, the misleading statements in the abstract, and to provide a more precise reflection of the methodology and the terminology used to explain it.

First, I suggest the title be revised. You cannot evaluate the performance of objective functions – they are means towards evaluation. It can be a comparison of different objective functions for assessing different regional climate models. And then the focus is clearly on Cyprus, not the eastern MED; otherwise, this is highly misleading.

In general, it would also be beneficial if the authors reflect on the additional value of not only assessing model performance compared to the past, but also conducting a sensitivity analysis. For example, Wagener et al. (2022) (https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcc.772) explain how response-based evaluation can be a complementary strategy. Connected to this matter is the lack of a sensitivity analysis to prepare for calibration. Without knowledge of what parameters can and should be changed, there is limited value in comparing the calibration to different metrics. The calibration also requires the authors to state which parameters are changed clearly and the corresponding value ranges. This also requires providing evidence on what parameters should be changed in the first place, i.e., a sensitivity analysis.

In response to the reviewer’s suggestions, we expand our model evaluation framework to include an analysis of the parameter values obtained under different objective functions, calibration periods (average, wet, and dry), and across the different watersheds. This analysis and a new figure help clarify how sensitive the model parameterization is to the choice of objective function and hydroclimatic conditions. This additional analysis is presented in the corresponding sections below.

We also change the manuscript title, from:
*“Evaluating the Performance of Objective Functions and Regional Climate Models for Hydrologic Climate Change Impact Studies: A Case Study in the Eastern Mediterranean”*
to
*“Evaluating Objective Functions and Regional Climate Models for Hydrologic Climate Change Impact Studies: A Case Study in the Eastern Mediterranean Island of Cyprus”*

### Specific comments

12: Compromised is the wrong word here. First, what does robustness mean here? Is it the ability to simulate with a similar skill under different circumstances? If that is the case, non-stationarity is something climate change is causing and might be a challenge, but it is not something that undermines model performance; rather, it questions whether models are fit for the right purpose.

Thank you. We will modify the sentence as follows:

“The use of hydrological models for projecting future freshwater resources is challenged by non-stationary climate conditions, as these conditions may affect whether models calibrated under historical climates are fit for use under future scenarios.”

16: But why would you evaluate objective functions? To check whether your calibration is good? However, that does not test the objective function; it tests how well your model was calibrated using different objective functions.

We will change this sentence to the following: “A comparative scheme was developed to investigate how the hydrological model performs when calibrated using six different objective functions.”

21: This is a misleading statement; you are simulating catchments in Cyprus, not in the MED. Similarly, your conclusion in line 30 is highly misleading as well.

We will correct the reference of “Mediterranean watersheds” in L21 to “the studied watersheds”.

40: Kang Ji 2023 Reference missing

Thank you. We will correct the citation from “Kang Ji et al. (2023)” to “Ji et al. (2023)”, and we will add the missing reference:

Ji, H. K., Mirzaei, M., Lai, S. H., Dehghani, A., & Dehghani, A. (2023). The robustness of conceptual rainfall-runoff modelling under climate variability–A review. Journal of Hydrology, 621, 129666. https://doi.org/10.1016/j.jhydrol.2023.129666

55: This seems like a terrible idea. Why would you restrict model simulations to such a subjective space? Unfortunately, I am not able to find the publication in the references to understand this in more detail.

We will clarify and correct L55 and move the revised sentence after the sentence in L46, as follows:

The review study of Ji et al. (2023) showed that, for credible model transferability between calibration and validation, the reported range for precipitation change towards drying conditions, by different studies, is narrower (i.e., from -10% to -30%) than the corresponding range for wetter conditions (i.e., from +10% to +80%).

56: Could you reflect on multi-objective function calibration here as well? Does this solve some of the problems, and if not, why not?

We will add the following sentence in L57:

“A composite objective function, comprised of a linear combination of different functions accounting for different aspects of flow regimes (Zhang et al. 2008), was tested against commonly used objective functions by both Fowler et al. (2018) and Munoz-Castro et al. (2023). Fowler et al. (2018) found that the multi-objective function was outperformed by simpler formulations, such as the Refined Index of Agreement (Wilmott et al., 2011), which uses the sum of absolute errors, and Kling-Gupta Efficiency (KGE; Kling et al., 2012), computed individually per year. According to these authors, these two functions also performed better than models calibrated on squared-error based measures.”

Reference: Zhang, L., Potter, N., Hickel, K., Zhang, Y., & Shao, Q. (2008). Water balance modeling over variable time scales based on the Budyko framework–Model development and testing. Journal of Hydrology, 360(1-4), 117-131. https://doi.org/10.1016/j.jhydrol.2008.07.021

81: Please specify how the projections differ for the different RCPs with respect to the study ranges you cite here.

We will modify the sentence in L81 as follows:

"CMIP5 climate projections under different Representative Concentration Pathways (RCPs) for the entire Mediterranean show temperature increases of 1.1, 2.2 and 4.4 °C under RCP2.6, RCP4.5 and RCP8.5, respectively. Precipitation changes are projected to be +1%, –6.6% and –18.8% for these scenarios over the long-term period (2081–2100), relative to 1986–2005 (Gutiérrez et al. 2021). With high-resolution (0.11° ≈ 12km) RCM simulations for the Mediterranean under RCP8.5, Zittis et al. (2021a) highlighted an annual precipitation reduction of up to 10% for the first half of the 21st century and reductions of up to 20%-40% for the second half, particularly for the southern and eastern areas."

New reference:

Gutiérrez, J.M., R.G. Jones, G.T. Narisma, L.M. Alves, M. Amjad, I.V. Gorodetskaya, M. Grose, N.A.B. Klutse, S. Krakovska, J. Li, D. Martínez-Castro, L.O. Mearns, S.H. Mernild, T. Ngo-Duc, B. van den Hurk, and J.-H. Yoon, 2021: Atlas. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, et al. (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1927–2058. 10.1017/9781009157896.021

91: Performance limits in what regard?

We will remove "performance limits" and modify sentence in L91 as follows:

"Assessing the impact of climate change on fresh water resources through hydrological modelling requires understanding model performance under a range of climate-change signals and magnitudes."

96: Again, you cannot evaluate the performance of an objective function. You can evaluate the model performance of a model that has been calibrated with a specific objective function or a set of objective functions.

We will change the specific terminology and replace it accordingly in the entire manuscript. Some indicative examples follow:

In Ln96: The specific objectives are: (i) to develop a method for evaluating the performance of ~~different objective functions for~~ a hydrological model ~~calibration~~ when calibrated with different objective functions under a changing climate.

Ln 246: "The comparison of the model performances achieved with ~~matrix of evaluation of~~ the six objective functions, based on the median values of the 38 watersheds (averaged for all 5-year periods) for each performance measure is shown in Table 1."

Ln 258: In the validation, the NSEsqrt-calibrated model also outperformed the other ~~objective functions~~ calibrated models in 14 watersheds, whereas the KGE-calibrated model ~~parameterization~~ outperformed the other ~~objective functions~~ calibrated models in three watersheds only, which is the second worst performance out of the six ~~functions~~ optimizations.

107: Why are you using these specific metrics and their transformations? What kind of behavior space are you covering with them? Why are you not also separating them into components that would specifically tell us something about the mass balance, peak behavior, etc.? And wouldn't this be more valuable to assess in a sensitivity analysis? This would also tell us which parameters are sensitive under which objective functions for different catchments in your assessment.

Thank you for your comments.

Concerning the selection of the specific metrics, we will add the following justification in Ln107:

“NSE and KGE were selected as the main objective functions due to their widespread use in hydrological modelling in both original and modified forms (Fowler et al. 2018; Guo et al. 2020). The use of the original, square-root, and logarithmic transformations ensures representation of high-, medium-, and low-flow conditions (Moriasi et al. 2007; Seiller et al. 2017). Percent bias (PBIAS) was also included to provide an independent measure of total volume error (Moriasi et al. 2007; Coron et al. 2012).”

In order to highlight the importance of conducting a sensitivity analysis of model components that might affect the results of the impact study we will add the following sentence in introduction in L55:

“A sensitivity analysis of the model response to multiple inputs provides useful insight into model behaviour under changing conditions, as recommended by Wagener et al. (2022).”

We will also add a model parameter analysis that shows the parameter value ranges under different objective functions, different calibration periods and for different watersheds. The results of this analysis will be added to the end of Section 3.1 “Selection of objective function under changing climate conditions”, as follows:

“The optimized model parameters derived using different objective functions and calibration periods were plotted against the average precipitation of the respective calibration periods (Figure A1). The results show that, for a given watershed, parameter values generally vary more between different objective functions than between different calibration periods. This finding suggests that understanding how different objective functions lead to different model outputs, in relation to the input model parameters and interannual variations in precipitation forcing, is important.”

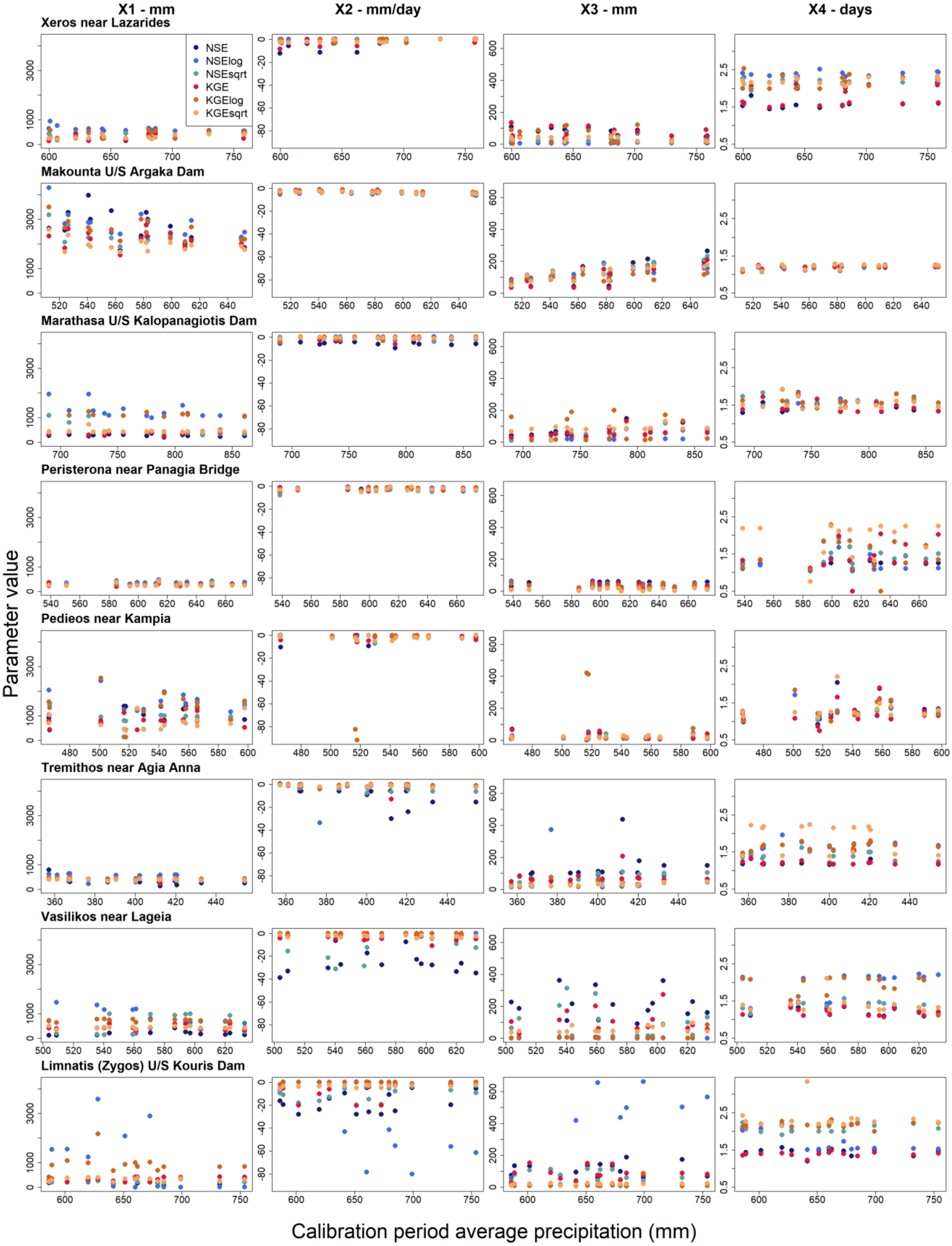

Figure A1: Values of the four parameters of the GR4J model optimized with six objective functions for the 15 five-year calibration periods for eight watersheds. The mean precipitation of each five-year calibration period is shown on the horizontal axis.

New reference:

Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. Wiley Interdisciplinary Reviews: Climate Change, 13(3), e772. https://doi.org/10.1002/wcc.772

135: How sensitive is the calibration to picking this specific value? How does this assumption impact the results?

We will add the following explanation in L136:

“The minimum temperature increase used to select calibration and validation runs was 0.7°C. This value did not exclude any five-year period from being tested for calibration, because every five-year window within the calibration period (1980–1997) could be matched with at least one five-year window in the validation period (1998–2015) that was warmer by 0.7°C or more.”