Response to Reviewer 2

We thank the referee for the helpful review and the constructive comments. His/her comments are given below in black, and our responses are in blue. Proposed changes to the manuscript are in red.

This manuscript investigated the ability of parameter transfer to ungauged basins in global hydrological models based on the calibration experiments of WaterGAP3. The combination of traditional and novel regionalization approaches revealed that machine learning-based methods may be too flexible for regionalizing. However, some findings are not convincing and need more discussions or elaboration, some terms need better definitions, and the figures need improvements. Therefore, I recommend a major revision. Please see below for my detailed comments.

Thanks for this comprehensive summary of the critical points. In summary, we restructured Chapter 3 to elaborate more intensely on the findings. By this, we added a KGE evaluation for one representative split-sample test (s. Chapter 3.3) and modified the figures (in Chapters 3.1-3.3) to show more precisely the differences between the methods. Moreover, we extended the analysis of the impacts on the runoff simulations in Chapter 3.4 to (1) gain a deeper understanding of the impacts of regionalization methods in run-off simulations and (2) be more in line with the improved title.

Further, we reduced the set of examined basins to create more valid results. This decision was based on the observation that using only the bias in monthly discharge is inadequate to define a “sufficient model performance” because it does not account for differences in timing or variance. Therefore, we added a minimal KGE in monthly discharge of 0.4 as additional criterion for selecting basins where WaterGAP3 is capable of representing well the hydrological processes (s. ll. 134-139).

Major comments:

1) Title. The term "Regionalization and its impact" is too vague to present the content of this manuscript. Elaborate the titles based on the novelty of the presented work.

Thanks for the suggestion. We adjusted the title to be more precise, highlighting that the study focuses on the regionalization of a GHM and evaluating its impact on runoff simulations. Also, we revised the text in several places to emphasize the novelty of the work, i.e., applying multiple regionalization methods for the first time on the global scale using WaterGAP3, evaluating the methods, and evaluating their impact on the runoff simulation.

Previous title: Regionalization and its impact on global runoff simulations: A case study using the global hydrological model WaterGAP3 (v 1.0.0)

New title: Regionalization *in global hydrological models* and its impact on runoff simulations: A case study using the global hydrological model WaterGAP3 (v 1.0.0)

2) The range of gamma values is too narrow. The cluster of values around the maximum range of gamma implies that the higher values seem appropriate but are excluded from the candidates. Please recalibrate gamma values using wider ranges, and present the scores in the accuracy of predicted streamflow in terms of score index (e.g., NSE).

Thanks for this suggestion. We understand that the parameter range and clustering of the calibrated parameter may be addressed more intensively to make it more comprehensible for peers. Therefore, we address this issue by a brief sensitivity analysis and discussion in the Appendix (see Appendix B in the revised manuscript). In a nutshell, high γ values are primarily sensitive when soil moisture is high and generally less sensitive than lower γ values. High γ values frequently occur in dry regions (see Figure 4b in Müller Schmied et al., 2021). In dry areas, it is not expected that the soil moisture has high values (e.g., see Khosa et al., 2020 and Oloruntoba et al., 2024 for estimated and measured soil moisture in Africa and Draper et al., 2008 for estimated and measured soil moisture in Australia). Therefore, it is not expected that higher gamma values will significantly enhance the calibration result or decrease the issue of clustered calibrated parameter values at the higher end of the parameter space. It is more likely that adding missing model processes, e.g., evaporation from rivers or inaccurate representation of groundwater processes, will solve the clustering of calibrated parameters for dry regions

(Eisner 2015, p. 49). Therefore, we do not change the parameter bounds for γ (please note that the same bounds are also used in Eisner 2015, p. 16; Müller Schmied et al., 2021; Müller Schmied et al., 2023).

As the model performance in calibration is critical when fitting a regionalization method to become valid, we sharpened the criteria for the basins used in this study. Thus, we now exclude all basins with a monthly KGE below 0.4 (933 basins remain after adding this additional criterion) (s. ll. 133-143).

3) The authors assessed the impact of regionalization only from the mean absolute errors in gamma values but neglecting the ability of parameters to represent the river discharges. The parameter transferred from one basin to another should reasonably represent river discharges (accepted behavior) at least in the donor basin. The success of the parameter transfer should be assessed by the accuracies of simulated river discharges in the transferred (ungauged) basins with the parameter. Missing information on the accuracies in the simulated river discharges (both in donor and transferred basins) impedes the reliability of this work.

Thanks for the valuable comment. We agree that including accuracy in simulated river discharge will increase the reliability of the paper. Therefore, we included a new chapter in the result section (see Chapter 3.3), where we added an evaluation of KGE values for one representative split sample.

As the model is very demanding in terms of computational time, conducting the analysis on all split-sample tests was not affordable. However, the additional analysis validates the assumption that logMAE values are a suitable tool for pre-selecting suitable methods for WaterGAP3. Nevertheless, we also recommend in the manuscript adding such an analysis when analyzing the regionalization of WaterGAP3, ”as the logMAE of calibrated and regionalized parameter values simplifies the inherent complexity between model parameters and model performance” (ll. 459-461). Please note that we modified the MAE to logMAE to account for the high sensitivity of smaller γ values.

Minor comments:

1) L 174. The usage of 'heavy-tailed' distribution seemed wrong. In probability distribution, heavy-tailed distributions have heavier tails than the exponential distribution. The authors may intend different meanings, but the use of 'heavy-tail' here would misleading.

Thanks for the hint. We modified the term and using know “cluster of calibrated parameter values at the extremes of the valid parameter space” or similar throughout the complete text (e.g., ll. 152, 333).

2) L102. Srmax values are not presented in the manuscript. Maybe it is based on the look-up tables of this model. Please provide it as many researches indicated that soil storage is an important parameter for GHM.

Thanks for the suggestion. We added a global map of the maximal soil moisture to the Appendix (see Appendix A in the revised manuscript) and supplied it in the GitHub repository as .tiff.

3) L259. Explain why 162 groups were used.

We have noted that the "highly flexible version" only indirectly satisfies our goal of analyzing if the clustering beforehand applying knn enhances the estimate. We have now implemented knn directly, where each basin refers to one "cluster", meaning that we used a knn-algorithm that defines the most similar donor basin for each pseudo-ungauged basin (based on Euclidean distance between min-max-normaliued attributes). We have modified the text accordingly (ll. 297-301) and adopted all figures in the restructured result section (s. Chapter 3) to evaluate the effect of using knn without clustering beforehand.