We thank the editor for the helpful review and the constructive comments. His comments are given below in bold, and our responses are given in italic. Proposed changes to the manuscript are in red. The information about the modified lines is referring to the new version of the manuscript without tracked changes. Additional references are provided at the end of this document.

Public justification (visible to the public if the article is accepted and published):
There are a few comments to be addressed.

In applying the machine learning techniques, could you construct feature importance score bar graph for insights on the influences of the various predictors on the targeted variable or parameter?

Thanks for suggesting importance score bars to give insights into the predictors' importance. As using the "p+cl" descriptor set for the tuned MLR version and knn results in the lowest logMAE values but is rather optimal (because it was initially considered as a control group for the selected subset of the descriptors), we use the feature importance bars to give first insights on how to optimize the descriptor sets for these two methods in further studies. The two feature importance score bars can be found in Appendix E (ll. 833-839).

The use of correlated predictors for the regionalization method comprises the influence of multicollinearity on the study results. There are many ways to avoid multicollinearity in modelling or regionalization. For instance, only one of any two significantly correlated predictors can be considered for modelling or analysis. Therefore, include some discussion on the possible future direction for addressing the issue of multicollinearity in your model or regionalization approach. You may find relevant information from the publication via https://doi.org/10.1002/wrcr.20315 to guide you in the discussion.

Thanks for emphasizing the importance of multicollinearity. A new section,"Challenges and Future Directions", has been added to the manuscript, in which the issue of multicollinearity is discussed in greater detail (ll. 686-696). We suggest several methods for dealing with multicollinearity, namely the use of PCR and PLS to create predictors with low multicollinearity and the explicit checking for multicollinearity in predictor sets using VIF.

Provide reference for Equation 1. Provide reference for the Kling-Gupta-Efficiency (KGE). See article published via https://doi.org/10.1016/j.jhydrol.2009.08.003 for the citation. Furthermore, provide the formulae for KGE and logMAE.

The two equations and the related references were added to the manuscript (KGE: l. 137, ll. 146-148, logMAE: ll. 258-261). Thanks for drawing attention to this lack of completeness.

Many methods exist for evaluating model quality. In the discussion, highlight on the point that the choice of an objective function comprises a sub-source of calibration uncertainty. Relevant articles on this can be found via https://doi.org/10.5194/piahs-385-181-2024 and https://doi.org/10.5194/adgeo-5-89-2005 . The idea is that, the use of other methods for model performance evaluation could be recommended for analysis to take into account the influence of the choice of an objective function on modelling results for regionalization.

Thanks for emphasizing the role of benchmark selection for model performance evaluation. Two new paragraphs have been added to the end of Section 3.3 in order to elaborate further on this issue (ll. 517-537). Section 3.3 is primarily concerned with the evaluation of the KGE. The newly added paragraphs explain the rationale behind selecting the KGE, emphasizing that the choice is purpose-dependent. Moreover, we created an additional evaluation using a modified version of NSE (as suggested by Krause et al., 2005) and supplied it to the appendix (s. Appendix F, ll. 840-856). Adding this analysis highlights the inherent uncertainty in model evaluation due to the imperfectness of evaluation criteria. Furthermore, it substantiates the value of ensemble runs, which we utilize in Section 3.4, enabling the selection of multiple regionalization methods for the evaluation of their impact on runoff simulations.

It is important to elaborate on why many predictors were considered. In fact, a robust regionalization-related model would that with as few predictors as possible. The idea of using many predictors (regardless of the value that each one adds) tends to be a trick to enhance performance of regionalization-related model. This is a misleading procedure. Thus, in such a case, it is recommended to make use of an adjusted R-squared, for instance following Ezekiel (1930) or https://psycnet.apa.org/record/1931-02963-000 to evaluate performance of the regionalization procedure.

Thanks for addressing this issue. This statement primarily concerns selecting the descriptor set "p+cl" for MLR tuned and knn. The two other selected methods, SP and SI ensemble ("subset"), are unrelated to this issue as SP uses spatial distances as the metric, and the chosen SI approach uses only five predictors for regionalization. We created two new paragraphs in Section 3.2 to elaborate further on the descriptor selection for the two methods (ll. 422-438) as we consider this issue to be highly relevant for enhancing the study's reliability.

In summary: We strongly agree that the use of a reduced number of predictors (or, in general terms, a minimal number of degrees of freedom) should be a fundamental consideration in the construction of models to reduce the risk of over-parametrization, i.e., increasing the model's robustness. Choosing 12 descriptors for tuned MLR and knn, we argue the following.

1. *The models remain stable in model quality with no discernible decline from training to testing indicating their robustness. This is evidenced by using 50 % of the basins for training and testing, repsectivley, thus leaving a high proportion of basins for testing (cf. Table D1).*
2. *Using 12 descriptors is in the order of other studies, even if is in the upper tail (e.g., McIntyre et al., 2013, used three predictors; Beck et al., 2016, used eight predictors; Chaney et al., 2010, used 13 predictors).*
3. *The initial hypothesis was that the descriptor set "p+cl" would function as a control group to validate the suitability of the selected subset of descriptors. Consequently, the predictor selection for tuned MLR and knn using the descriptor set "p+cl" is not optimal, although it is robust. It would be beneficial to optimize this in future studies, for example, by considering feature importance scores.*

It should be noted that the use of R-square in place of logMAE would emphasize larger errors, namely those occurring in the upper (less sensitive) tail of the parameter range. Therefore, we omit using the adjusted R-square as an additional metric as it may yield erroneous implications in the context of our study.

Identify and elaborate on the general challenges and/or limitations of your study or regionalization. Also highlight on the future directions to guide on how the relevant improvements could be made.

Thanks for addressing this issue. A new Section 3.5, "Challenges & Future Directions", has been added to the manuscript in order to address this issue (ll. 662-696). In this chapter, we addressed the considerable runtime required for GHMs, which impedes the execution of extensive regionalization experiments, the predictor selection for regionalization, where no universally valid approach exists and where our study can further be enhanced, and the explicit consideration of multicollinearity in the case of our study. Moreover, we added some potential future directions into the conclusion section (ll. 713-715) and modified the last sentence in the abstract (ll. 28-29).

References:

Krause, P. Boyle, D. P., Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Advances in Geosciences, 5, 89-97, https://doi.org/10.5194/adgeo-5-89-2005, 2005.