Comments on manuscript entitled ‘Multi-Machine Learning Ensemble Regionalization of Hydrological Parameters for Enhances Flood Prediction in Ungauged Mountainous Catchments’ by Li et al.

The manuscript deals with developing a multi-machine learning ensemble method for regionalization of a hydrologic model (Top-SSF) over 80 catchments in southwestern China. The authors showed the improvement in performance using multi-machine learning method over single methods. While the manuscript is well-structured and results are clearly presented, there are some points need to be addressed before the publication of the manuscript. Please find the comments as follows:

Response: Many thanks for your comments.

1) Line 107: what’s the range for catchments area?

Response: Added ‘ranging from 109 to 6564 km’ in Section2.1.

2) Legend of Figure 1: please use the term ‘Hydrometry station’

Response: Revised

3) Line 122: Hourly flow data

Response: Revised

4) Line 150: TOPMODEL not TOPMODE

Response: Revised

5) Section 3.1: More details should be provided. For example: What kind of hydrologic model is Top-SSF? Continuous or event-based? Lumped or (semi)distributed? And how it is going to be applied in this research? To simulate flood events? Or a whole time series (continuous modelling)? What are the inputs to the model, e.g. precipitation and temperature data?

Response: More details of the Top-SSF model have been added in Section 3.1 and Section 3.3.1 of the revised manuscript.

Top-SSF is a semi-distributed hydrological model based on the well-established TOPMODEL framework, which delineates sub-basins based on the topographic index. It retains the key advantages of TOPMODEL, such as its parsimonious structure, physical interpretability, and ease of parameter transfer. In this study, while the model was driven by the continuous hourly meteorological data (including precipitation, temperature, surface pressure, relative humidity, wind speed, and net solar radiation), it was applied in an event-based manner to specifically simulate flood events. For each catchment, the model was calibrated using two independent, representative flood events and validated against a third, distinct flood event.

6) Result section, Lines 362-365: Why performance of the different machine learning methods for parameter regionalization is compared against the Top-SSF model and not against the observed flood events?

Response: The experiment was intentionally designed to compare the results of the machine learning methods with that of the Top-SSF model, isolating the performance of the parameter regionalization method themselves. To clarify this rationale, we have added a supplementary discussion in Section 5.5 of the revised manuscript. This method is based on two primary justifications:

First, the fundamental aim of the parameter regionalization is to effectively transfer model parameters to ungauged catchments, not to reconstruct or alter the model's underlying structure. By using the calibrated Top-SSF simulation as the benchmark, the theoretical "best-case" performance for that specific model structure was established. Consequently, any performance degradation observed in the regionalized models can be directly and exclusively attributed to the defects of the regionalization method, rather than being confounded by the inherent structural limitations of the hydrological model.

Second, this method ensures that we are assessing the regionalization method's ability to learn the underlying model physics, not to mimic data noise. While the Top-SSF model is calibrated against observed data (which is subject to measurement uncertainty), its output is a structurally consistent representation based on its physical equations. If we were to use the raw observations as the target, the machine learning methods might derive "spurious" parameter sets that compensate for both the hydrological model's structural errors and the observational errors. Such parameters might appear effective but lack physical meaning and generalizability. By targeting the Top-SSF simulation, we force the ML methods to learn the intended relationship between catchment attributes and the model's parameters, leading to a more robust and physically interpretable assessment of the regionalization techniques.

7) Figures 11a and d: how can the NSE be greater than 1?

Response: **It is absolutely correct that the Nash–Sutcliffe Efficiency (NSE) has a theoretical upper bound of 1 which represents a perfect model simulation. However, in our initial visualization of Figures 11a and d, we used standard deviation to construct error bars for NSE, Qp, and Tp. It can be misleading for NSE due to its bounded nature (ranging from –∞ to 1). As a result, the error bars erroneously extended beyond the physical limit of NSE = 1. To address this issue and ensure the accuracy of uncertainty representation, we have revised the calculation method for the error bars of NSE. Instead of using standard deviation, we now use the range (i.e., the difference between the maximum and minimum values) across the donor catchment configurations under each scenario. This revision ensures that the error representation remains within the theoretical bounds of NSE while still reflecting the variability of the model performance across the different donor catchment selections.**

This correction improves the clarity and scientific validity of our results presentation without altering the main findings of the study.

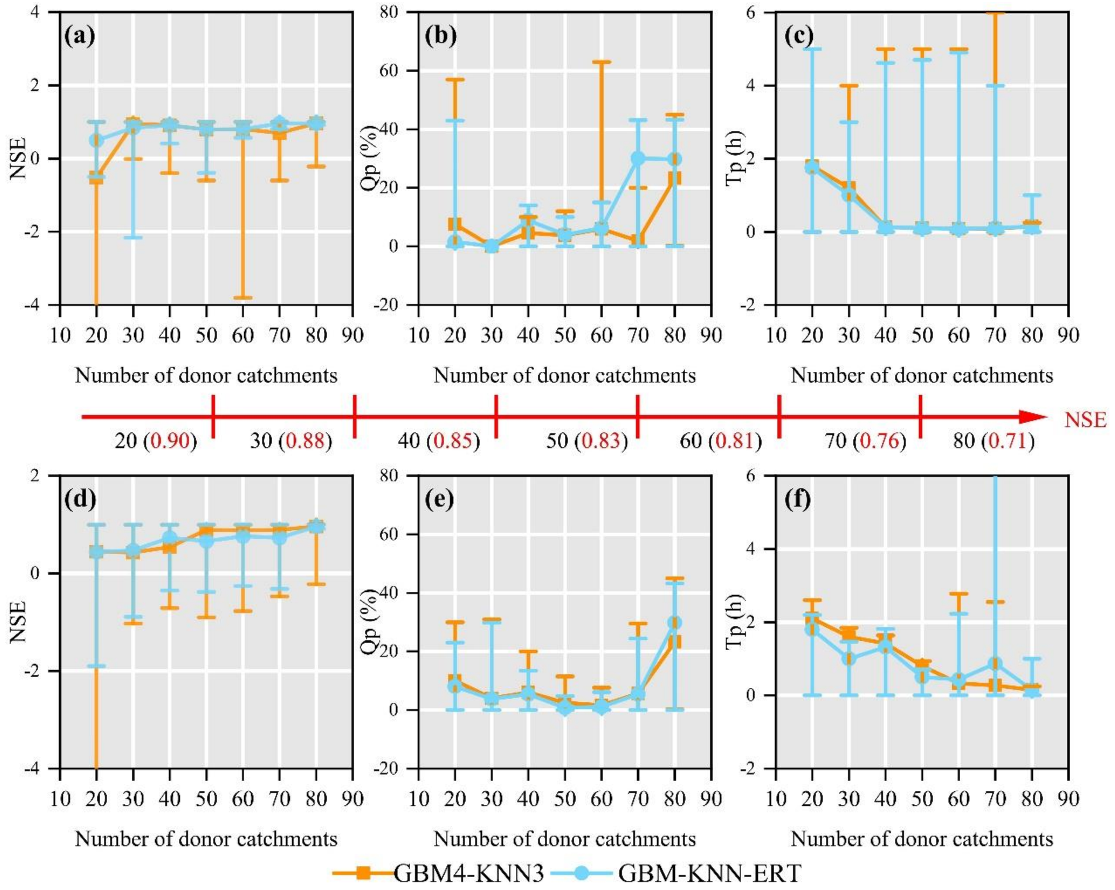

Fig. 11. Performance comparison of two donor catchment selection methods for parameter regionalization as a function of donor catchment quantity. Mode1 (a-c) selects donor catchments in order of decreasing NSE, while Mode 2 (d-f) selects them randomly. Flood prediction accuracy is assessed using NSE, Qp, and Tp. Error bars represent the full range (minimum to maximum) of the performance metrics.

8) Section 5.4: Not clear how the calculations carried out to simulate peak discharges. Which events in future are selected for this analysis? Did the whole time series of projected precipitation in baseline and future periods fed to the hydrologic model? Or just a few storms selected?

Response: Yes, this is not clear. Clarifications have been added in this section. Specifically, in this part of the study, the impact of climate change is reflected through the changes in two catchment descriptors, i.e., mean annual temperature (Tem) and mean annual precipitation (Prec). Specifically, for the historical period, Tem and Prec represent the multi-year averages over 1901–2021; while for the future period, they represent the projected multi-year averages over 2022–2100 under the SSP5-8.5 scenario. To assess the influence of these climatic changes on flood prediction performance, we applied the parameter regionalization models (GBM4-KNN3 and GBM-KNN-ERT) calibrated by using historical data to future conditions. Under the unchanged model structures and hyperparameters, only the historical Tem and Prec values were replaced with their corresponding future projections. The simulated peak discharges were derived from the three flood

events used in the calibration and validation of the Top-SSF model. We then compared the maximum flood peak discharge across all simulations between the historical and future periods to evaluate the absolute differences in runoff modulus.

This approach allowed us to isolate the effects of projected climate change on the stability and robustness of the parameter regionalization methods, particularly focusing on how changes in temperature and precipitation patterns influence flood peak predictions in the ungauged mountainous catchments.