This review was prepared as part of graduate program course work at Wageningen University, and has been produced under supervision of Ryan Teuling. The review has been posted because of its good quality, and likely usefulness to the authors and editor. This review was not solicited by the journal.

REVIEW of the paper "Improving Runoff Simulation in the Western United States with Noah-MP and VIC" by Lu Su et al.

This manuscript studies streamflow forecasts improvement in the Western U.S. using VIC and the Noah-MP model, also evident in the title. The authors describe a systematic calibration of parameters for VIC and Noah-MP resulting in model accuracy improvement. The calibrated parameters were extended to ungauged basins and the entire region using the donor-basin regionalization method. Both models showed improvement in the high and low flow simulation capabilities after calibration and regionalization. The structure and organization of the paper is coherent. The study uses suitable models and perform highly actionable simulation. Developing parameter sets regionalization across all HUC-10 basins in the WUS seems relatively novel. The topic of this work is of interest to the regional water management practitioners. The study is valuable for the regional streamflow simulation and prediction in the Western United States. It fits the scope of the journal, very relevant for HESS. The manuscript has a clear potential for publication, though there are a few aspects that need to be clarified. Based on my comments below, I recommend moderate revision before the manuscript can be published.

Major arguments

Two models used: The authors need to provide a better explanation of why they decided to use two models separately to improve runoff simulation. In fact, there are two hydrological models applied in previous studies. Two models are selected as representatives of different levels of model complexity to see how model complexity differences impact findings, and are used to provide a reliable empirical assessment in the experiment (Shen et al., 2022). Overall, this manuscript may not explicitly provide a direct comparison of the forecast results for the VIC and Noah-MP models. Though possible explanations on VIC outperformed Noah-MP both pre- and postcalibration are given, and there is a quick mention of the regionalization enhancement greater for the Noah-MP model compared to the VIC model. The limited comparison is more like explaining the results by corresponding to the previous text. Instead of just stating that both the VIC model and the Noah-MP model are used for streamflow simulation improvement, there should be a satisfying reason to

use two models. It is not clear why you study both models when they each could improve the simulation accuracy. I believe the authors should distinguish at least slightly between the two models used. Perhaps there could be some discussion between the two models, which model works better in which situation. Or, you suggest what to do with the two results to forecast.

Thank you for your feedback on our manuscript. We appreciate this opportunity to clarify our selection of the VIC and Noah-MP models for our hydrological study in the Western United States (WUS). As indicated in our response to reviewer 2, both of the two models we used (Noah-MP and VIC) are recognized for their effectiveness in hydrological studies, as evidenced by their widespread use in both the U.S. and globally (Mendoza et al., 2015; Tangdamrongsub, 2023). VIC's reputation for effective runoff simulation and Noah-MP's emerging role as the hydrological core of the National Water Model (NWM) make them ideal choices for our research objectives.

A previous comparative study of LSMs in the CONUS using the NLDAS test bed highlights the strengths of these models - Noah-MP provides the best performance in simulating soil moisture and is among the best in simulating total water storage, and VIC ranks the highest in performing the streamflow simulations (Cai et al.,2014). This underscores their complementary capabilities.

In light of your feedback, we will enhance the manuscript by: (a) providing a more detailed explanation of the rationale behind choosing both the VIC and Noah-MP models, focusing on their distinct physical parameterizations and their relevance to our research objectives; (b) Elaborating on how the contrasting approaches of these models help in addressing the variability and uncertainty in hydrological modeling, thereby providing a more robust analysis.

In choosing Noah-MP and VIC, we aimed to leverage their unique physics methodologies, crucial for addressing the diverse hydrological scenarios of the WUS. This approach allows us to encompass a wide range of hydrological behaviors and their spatial variations, providing a more robust and comprehensive hydrological analysis.

We believe these details offer a clear rationale for our model selection and hope these explanations enhance the readers' understanding of our methodological approach and its significance in advancing hydrological studies.

Parameter sets:

The paper concludes that gridded parameter sets were developed for both the VIC and Noah-MP models to all 4816 HUC-10 basins across the WUS after calibration and regionalization. However, the process of obtaining the parameter set seems a bit vague, with few direct mentions

in the manuscript. The authors calibrated 6 parameters for VIC and 5 for Noah-MP. The next regionalization process requires basin-specific features taken into account, introducing more information from the ungauged basins. Will this result in the necessity of more parameters? Will free parameters be brought in? Perhaps the size of the parameter space could be clarified. The particular applications of the two models are also assumed to be different. Probably additional modifying parameters should be involved in the process to make the model transferable across space (Feigl et al., 2022). The gridded parameter sets could be further explained to indicate a centralized view. I think how the parameter sets are developed should be further discussed.

Thank you for your comments. We will give a clearer explanation of the parameter calibration and regionalization processes in the revised MS. In short, we calibrated six parameters for the VIC model and five for the Noah-MP model, focusing on those that significantly influence soil moisture and runoff simulation. Following the calibration process, we regionalized the parameters from gauged to ungauged basins based on a mathematical assessment of the spatial and physical proximity between the gauged and ungauged basins. Specifically, we employed a donor basin approach where parameters from a gauged (donor) basin were applied to an ungauged (recipient) basin based on their similarity. To determine this similarity, we employed 18 basin-specific features, primarily of geospatial and climatological nature, as detailed in Table S1. The parameters used for calibration and the features used to determine the similarity index in the regionalization process are under different categories. This regionalization won't result in more parameters or free parameters. The physics that control the key hydrological processes of the two models are different, so we explored their best regionalization features separately.

For each of the 4816 HUC-10 basins, we calculated a similarity index with the calibrated basins using the selected features. The top three most similar basins were identified as donor basins, and their weighted average parameters were then adopted by the target basin.

We acknowledge your suggestion regarding the potential development of a parameter transfer function. While not addressed in our current study, this may be a valuable direction for future research.

In response to your comments, we will revise our manuscript to include a more thorough explanation of the regionalization process.

Best regionalization features:

Selection of relevant catchment features is imperative for the success of regionalization (Bastola et al., 2008). It is not clear to me exactly how the best regionalization features are derived. The authors describe that the addition of further features doesn't improve KGE. It is not evident how you defined further features in the best regionalization features. Based on what is stated in line 360, it seems that each feature is added in a particular order. But the sequence is not specified. I think the authors should give more explanations on the applied iterative approach. In fact, relationships could be found between features. Therefore, these features could be fixed on the basis of the correlations, for example. Then the iterative process was employed by varying other features (Narbondo et al., 2020). I suggest the authors to be clearer on this point. Perhaps there could be a list indicating the importance of the features to give the rank. I would like to see more discussion here.

Thank you for your comment; we will add a more detailed explanation of this process in the revised MS.

To determine the most effective regionalization features from the 18 basin characteristics listed in Table S1, we employed a systematic iterative approach. the first iteration includes 18 simulations, each incorporating one of the 18 features. The feature that yielded the greatest increase in the median KGE across all basins, based on leave-one-out cross validation, was then retained. In the second iteration, we conducted 17 simulations, each combining the retained feature from the first iteration with one of the remaining 17 features. The feature that, in combination with the previously selected feature, resulted in the greatest further increase in median KGE, was then retained. This process was repeated iteratively, reducing the number of features considered in each subsequent round. The selection process continued until the addition of new features no longer resulted in an appreciable increase in median KGE. The sequence of the features shown in Figure 8 indicated the importance of the features and we'll make it clear in the revised manuscript.

This iterative approach ensured that each feature's individual and combined contribution to model performance was thoroughly assessed. It allowed us to identify a subset of features that, when used together, optimally improved model accuracy.

We acknowledge the possibility of correlations between features, which could influence their effectiveness when combined. We recognize the potential value in examining these correlations to further refine our feature selection process. In future studies, we intend to explore the relationships between features and how they can be leveraged to enhance the regionalization approach.

Minor arguments

The study only considers the KGE metric for model evaluation, which may not capture all aspects of streamflow simulation performance. The results could be supplemented by other evaluation metrics.

The decision to use KGE was based on its ability to comprehensively capture essential aspects of hydrological model performance, including correlation, bias, and variability between observed and simulated streamflows. KGE's widespread acceptance and utilization in hydrological simulation evaluations have been due to its effectiveness in providing a balanced assessment of these critical factors (Gupta et al., 2009; Konben et al., 2019). See also our response to reviewer 1; in the revised MS we will examine the effects of alternative (to KGE) objective functions.

Check your references. Some of the references are not shown in the references part even they are put in the main text. Please complete this section to provide sufficient details so that readers can locate the source of each citation.

Thank you for pointing out the inconsistencies in our references. We apologize for any missing citations in the reference section and we'll correct them in our revised manuscript.

Section 3.1: I suppose the obtained VIC model parameters seem to be too region specific. Perhaps indicate if the simulation can be replicated in a different area.

The parameters were calibrated to align with the unique physical characteristics of each basin. However, the calibration methodology we employed is broadly applicable and can be adapted to other regions. Other areas seeking to utilize the VIC model can apply this same calibration approach to identify their optimal parameters, ensuring a fit that reflects their unique hydrological contexts. Additionally, our regionalization method offers a pathway for transferring calibrated parameters from similar basins. This approach can facilitate the application of our study's findings to comparable regions, allowing for a more widespread utilization of the developed parameters.

p9, Table 1: The first column could probably have a better layout.

Thank you, we will make it better layout.

p10, line 213-216: This sentence might be split to express.

Thank you, we'll fix it.

p12: Perhaps move this paragraph forward, not to put up "3.2 Noah-MP parameterization" alone.

Thank you, we'll fix it.

p14, line 169: It might be good to have a reference for such a statement.

We'll add the references Livneh et al (2013) and Su et al (2021) .

p16, Figure 4: The figure name can be shown in full, adding (3), (6).

Thank you for pointing it out. We'll fix it.

p20, line 341: Please consider regionalization performance doesn't show significant increase when using more than 4 catchment descriptors to compute the Similarity Index (Poissant et al., 2017).

Thank you for your comment. The best feature combinations and numbers might differ for different models and in different regions. We explored it specifically for both VIC and Noah-MP and determined their separate sets of best regionalization features. Thank you for referencing the findings of Poissant et al. (2017) regarding regionalization performance and the use of catchment descriptors. We recognize the importance of this observation.

We tailored our approach to the specific requirements of both the VIC and Noah-MP models. We agree that the optimal number and combination of catchment descriptors for computing the Similarity Index can vary based on the model and the region. Therefore, we conducted a detailed exploration to identify the most effective set of descriptors for each model in our study area as shown in section 4: regionalization.

p22, line 373: The references do not have the evidence that geographical similarities are most significant (Burn and Boorman, 1993). Perhaps remove "This suggests that geographical similarities are the most important factor in parameter information transfer from gauged to ungauged basins."

Thank you for pointing this out. We'll edit as you suggest in our revised manuscript.

Section 6: The limitations of the study should be emphasized in the discussion section.

Thank you for your comment. We'll add more discussion on the limitations of our study. This includes the two different calibration method employed for VIC and Noah-MP, the potential limitation of regionalization method, the applicability of our findings to other hydrological contexts, and any limitations associated with the data used. We believe that openly discussing these limitations will not only provide a more complete picture of our study but also guide future research in this area.

References :

Bastola, S., Ishidaira, H., & Takeuchi, K. (2008). Regionalisation of hydrological model parameters under parameter uncertainty: a case study involving topmodel and basins across the globe. Journal of Hydrology, 357(3-4), 188–206. https://doi.org/10.1016/j.jhydrol.2008.05.007

Feigl, M., Thober, S., Schweppe, R., Herrnegger, M., Samaniego, L., & Schulz, K. (2022). Automatic regionalization of model parameters for hydrological models. Water Resources Research, 58(12). https://doi.org/10.1029/2022WR031966

Narbondo S, Gorgoglione A, Crisci M, Chreties C. Enhancing Physical Similarity Approach to Predict Runoff in Ungauged Watersheds in Sub-Tropical Regions. Water. 2020; 12(2):528. https://doi.org/10.3390/w12020528

Shen, H., Tolson, B. A., & Mai, J. (2022). Time to update the split-sample approach in hydrological model calibration. Water Resources Research, 58(3). https://doi.org/10.1029/2021WR031523

References:

Gupta, H. V., et al.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377, 80-91,2009.

Knoben, W.J., Freer ,J.E., Woods, R.A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences. 25;23(10):4323-31,2019 Oct.

Mendoza, P.A., Clark, M.P., Mizukami, N., Newman, A.J., Barlage, M., Gutmann, E.D., Rasmussen, R.M., Rajagopalan, B., Brekke, L.D. and Arnold, J.R.: Effects of hydrologic model choice and calibration on the portrayal of climate change impacts. Journal of Hydrometeorology, 16(2), 762-780, 2015.

Tangdamrongsub, N.: Comparative Analysis of Global Terrestrial Water Storage Simulations: Assessing CABLE, Noah-MP, PCR-GLOBWB, and GLDAS Performances during the GRACE and GRACE-FO Era. Water, 15(13), p.2456, 2023.