1. Response to editor

08 Jan 2024

Editor decision: Reconsider after major revisions (further review by editor and referees)

by Markus Hrachowitz

Public justification (visible to the public if the article is accepted and published): Dear authors,

As you have seen we have received detailed and very constructive assessments of your manuscript from two reviewers and one additional community comment. Overall, the objective and intention of your work has been positively received. However, the reviewers also flag a number of points that need to be adequately addressed before the manuscript can be considered for publication. From my perspective, in particular the following points will warrant some more attention:

Dear Editor, Thank you for your comments, and those of the reviewers. We have made the revisions as suggested and have addressed all the comments in the revised manuscript.

(1) As pointed out by both reviewers, the manuscript reads more like a technical report than a scientific paper. This aspect has to be substantially developed. In particular, it remains unclear what the novelty of the analysis is. In addition, it will be important to considerably expand on what can be learned from (or not) your experiment and what the implications thereof are. In the end, one of the objectives of a well-developed scientific paper is to pro-actively avoid the reader wanting to ask: "So what?"

Thank you for your comments. We have undertaken significant revisions. Specifically, we have enriched the introduction to better outline the research objectives and the novel contributions of our study. We have added a discussion section to delve into the insights and broader implications of our research and limitations, and how we can advance in future studies. Furthermore, we have restructured the conclusions to succinctly summarize the key findings and their significance. These enhancements aim to improve the overall coherence and impact of our paper, making the research goals, novelties, and implications clear to the reader.

(2) The rationale behind the choice of using these two models (and not others) will benefit from being strengthened as well.

Thank you for your suggestion. To address your comment as to the rationale for selecting the two models, we have enhanced section 2.2, specifically between lines 150 and 224. In this revised section, we provide a more detailed explanation of our choice, highlighting the unique attributes and relevance of these models to our research objectives.

(3) I also agree with the reviewers that the choice to limit calibration efforts to time series of stream flow and one single objective function comes a bit surprising in the light of the increasing body of literature over the past one or two decades that illustrates the benefit and even the necessity to develop much broader model calibration/evaluation schemes, either using multiple different objective functions (including spatial pattern) and/or calibration variables (e.g. hydrological signatures or other observable variables, e.g. snow cover).

Thank you for your valuable feedback. We recognize the critical role of employing a diverse set of metrics for a holistic evaluation of hydrological models

like VIC and Noah-MP. In response to this and Reviewer 1's comments, we have incorporated additional explanations in the manuscript, particularly at lines 349-355, 520-554 and 580-588. Our research predominantly focused on the calibration of streamflow, as it provides a comprehensive representation of catchment hydrology dynamics. Streamflow is a key term in the water balance. Given that we force with precipitation, in the long term mean reproducing streamflow implies that we're reproducing evapotranspiration, hence latent heat. We selected daily streamflow calibration as the primary focus because of its extensive applicability in the field of hydrology. Besides, observed streamflow is (much) more available and widespread across our domain compared to alternative metrics such as soil moisture or evapotranspiration.

Furthermore, we evaluated both models' representations of snow processes. Before we conducted the calibration, we conducted snow simulation verification at 20 selected SNOTEL sites across WUS. Our assessment indicated that the existing parameterizations for snow processes in both models were adequate for our study region. We've revised our manuscript in the calibration section at lines 258-267 to clarify these points.

To provide a more comprehensive analysis, we have included additional performance measures including NSE and BIAS, in the appendix of our revised manuscript (Figures S1 and S2). These measures complement KGE and facilitate a more comprehensive evaluation of model performance. We have also expanded our discussion on this topic in the manuscript, specifically at lines 349-355 and 580-588, to clarify our methodological choices and the rationale behind them.

(4) it remains completely unclear what the benefit of running the model at a sub-daily time-scale is, when the input data are daily. How can this be meaningfully done?

What is the benefit? How are modelling artifacts due to averaging effects in the daily data avoided at the sub-daily time scale?

We have provided additional clarification within the manuscript, specifically between lines 244 and 249. In short, Noah-MP requires a sub-daily time step, even if the output are aggregated (as we have done) to daily.

On balance, this manuscript can have some potential. However the above points and the remaining reviewer comments need to be addressed in detail and convincingly, which may require some substantial re-thinking and re-working of the experiment. However, I believe this could be possible in a round of quite major revisions.

I am looking forward to receiving a revised version of your manuscript.

Best regards,

Markus Hrachowitz

2. Response to reviewer 1

I enjoyed reading this fruitful manuscript. I only have several concerns and comments listed below.

-L186: the readers would except a similar treatment of the models for consistency of the framework. Except for the objective function (KGE), all other experimental details are different. Different search algorithms (SCE vs DDS), different max iterations (may be understandable). The chaotic structure of the paper makes is difficult to enjoy the fruitful outcomes. Currently it reads like a technical report by the corps engineers doing everything for better KGE. However, the research design needs a framework. The experiments should focus on effects of one thing one at a time. Using two different models is important for structural uncertainty; however under the same P, PET, Tavg input and methodology (calibration algorithm, max iterations etc).

Thank you for your comment. Our selection of calibration algorithms was guided by the need to balance computational efficiency with robustness. In the calibration of the VIC model, we chose the Shuffled Complex Evolution (SCE-UA) algorithm, a method well-established and widely recognized for its efficacy with this particular model (and with which we have considerable experience). The SCE-UA has been a benchmark in calibrating VIC for decades (see Naeini et al., 2019). The computational efficiency of VIC made SCE-UA a suitable choice, despite its requirement for a higher number of iterations. In practical terms, iterating a 20-year simulation in VIC takes about 2 minutes for a mid-sized basin, which we found manageable in terms of the computer resources available to us.

On the other hand, the Noah-MP model is more computationally demanding, and required a different approach. For this reason, we selected the Dynamically Dimensioned Search (DDS) algorithm. DDS is also used in the CONUS implementation of the National Water Model, which uses Noah-MP as its hydrologic core (Gochis et al. 2019). Although we had not used DDS previously, the fact that we had available to us a computational structure which embedded Noah-MP, in addition to its computational efficiency, was a deciding factor. We found that the DDS algorithm achieves optimal calibration with fewer iterations compared to SCE-UA (about 3000 iterations to reach optimal results for SCE-UA vs only about 250 iterations for DDS).

To validate our approach, we compared SCE-UA and DDS in calibrating VIC across 20 randomly selected basins. The results showed similar performance, reinforcing our decision to employ different algorithms suited to each model's computational needs. We could have run all our basins with DDS, but given the similarity of results for the selected basins, we didn't see any need (we note that we performed calibration on all of the basins for VIC before starting with the Noah-MP calibration).

While using the same calibration method for both models could simplify comparisons, as noted above, the two methods produce essentially the same results, the only difference being computational efficiency. We included a detailed explanation in our revised manuscript at lines 279-297 to better explain our methodological choices and their rationale.

Please a conceptual diagram/framework describing your methodology.

Thanks for the suggestion; we now include such a diagram in Figure 1 in our revised manuscript (reproduced below). This diagram illustrates our calibration

framework, delineating the distinct but complementary approaches we took for each model.

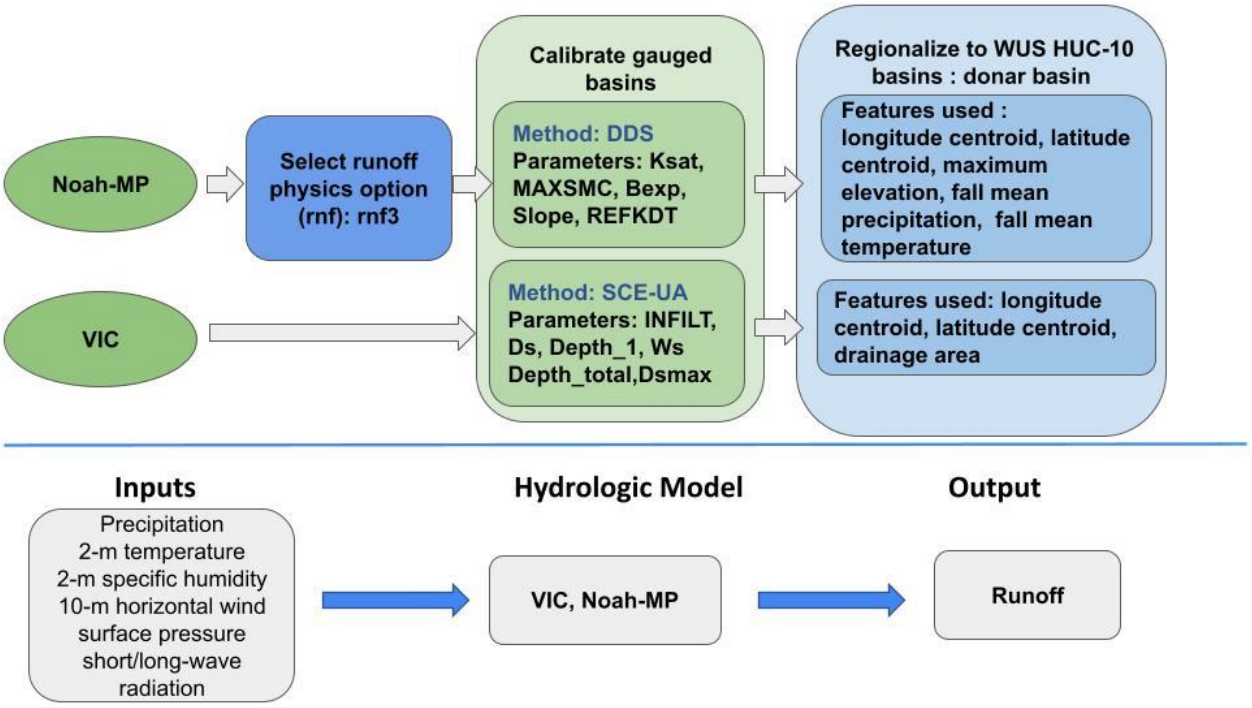

Figure 1 (a) framework of the calibration and regionalization processes adopted in this study. (b) model simulation inputs and output.

-Metric: Both models are top quality spatially semi or fully distributed models. Using a metric focusing on average flows may not fit best while NSE focuses more on high flows. There are other metrics focusing on patterns of simulated flux maps such as SSIM, FSS, EOF, SPAEF and SPEM. For novelty, it is recommended to elaborate calibration approach for the two models.

Thank you for your suggestion. We opted for the Kling-Gupta Efficiency (KGE) metric for daily streamflow evaluation, as it is a widely recognized performance measure that effectively considers bias, correlation, and variability (Gupta et al., 2009; Knoben et al., 2019). While we acknowledge that KGE provides a balanced assessment, we understand the potential benefits of other metrics like SSIM, FSS, EOF, SPAEF, and SPEM, especially for analyzing patterns in simulated flux maps. We

will consider some of these additional metrics in our future work. Specifically, we include (in the appendix) plots (Figures S1&2) of multiple performance measures (NSE and BIAS), which will allow direct comparisons of other measures with KGE and added sentences at lines 349-355 and 579-584.

Furthermore, while we used daily KGE as the objective function, which aligns with the general focus of our study, we evaluated the models' performance in predicting both high and low flows. This evaluation demonstrated that both VIC and Noah-MP models are proficient in high and low flow prediction, showing significant improvements post-calibration and after regionalization. We have expanded slightly our discussion of high vs low flow performance at lines 520-554 and 584-588.

-Did authors apply a sensitivity analysis before the calibration to select most important parameters for the streamflow using selected metric, KGE?

Prior to calibration, we conducted a sensitivity analysis to identify the most influential parameters for streamflow simulation, aligning with our selected metric, KGE. We have incorporated additional sentences about this analysis in Section 3.1 (lines 268-278) of our revised manuscript. Drawing on insights from previous research, we initially identified a comprehensive set of parameters. We then performed a sensitivity analysis, focusing on how variations in these parameters impacted KGE outcomes. This analysis allowed us to ascertain the parameters with the most significant impact on streamflow simulations. Considering the results of this sensitivity analysis and our available computational resources, we chose to calibrate six parameters for the VIC model and five for the Noah-MP model. This decision was made to ensure an efficient yet effective calibration process, balancing the need for accuracy with computational feasibility.

-Figure 4: For process consistency, different remote sensing products could be used to evaluate model results. Why did you only focus on streamflows, if the models are capable of producing AET, SM (at different soil horizons). MODIS, SMAP, SMOS, ESA, ALEXI all are useful constraints for model calibration making the novelty maximized. Literature review of the paper misses all those calibration papers focusing not only KGE but also RS products.

https://doi.org/10.1002/2017WR021346

https://doi.org/10.5194/hess-22-1299-2018

Thank you for your suggestion. Our current study primarily focused on streamflow calibration due to its comprehensive reflection of catchment hydrology. However, we recognize the potential of remote sensing products like MODIS, SMAP, SMOS, ESA, and ALEXI in providing additional data for calibration, particularly for variables such as actual evapotranspiration (AET) and soil moisture (SM). While such remote sensing products can be valuable for calibrating and validating hydrological models, our current study was limited by the availability of observed soil moisture and evapotranspiration data.

Nonetheless, we understand the importance of these data sources in enhancing model calibration and validation. In future studies, we aim to incorporate additional data sources to calibrate and validate other hydrological variables, enriching the scope and accuracy of our models. The references you provided will be a valuable addition to our literature review, guiding us to include a broader range of calibration variables. We have revised the manuscript to comment on this broader perspective and acknowledge the importance of integrating remote sensing data in hydrological modeling at lines 77-81, 258-262, 568-578 .

-Do these models incorporate pedo-transfer functions for parameter regionalization? Some of the distributed models, like mHM, uses multi-parameter regionalization (MPR) technology.

https://doi.org/10.5194/gmd-15-859-2022

No, we did not utilize pedo-transfer functions for parameter regionalization. Our focus was on direct calibration of model parameters for each basin individually. Our method involved calibrating parameters specific to each basin, taking into account their unique hydrological characteristics. Following this, we transferred these calibrated parameters to HUC10 basins based on similarity assessments. This approach (Bass et al.,2023) ensured that the calibration was closely aligned with the specific conditions of each catchment area. We acknowledge that alternative approaches, such as the multi-parameter regionalization (MPR) technology used in models like mHM (referenced in your citation), provide different perspectives on parameter regionalization. We appreciate your comment and will consider alternative regionalization methodologies, including pedo-transfer functions, in our future research to enhance the depth and applicability of our work. We comment on this possibility at lines 588-596.

-Table 1: “VIC4.1.2”

Why this old version of VIC model is used while current version WRF-Hydro 5.2.0 is preferred.

VIC5 version includes many infrastructure improvements (glaciers etc) as described here:

https://doi.org/10.5194/gmd-11-3481-2018

We selected VIC 4.1.2 for two key reasons: Firstly, our initial parameters were based on Livneh et al. (2013), who validated model discharges over major CONUS

river basins using VIC 4.1.2. To leverage these well-established parameters and maintain consistency with their study, it was crucial to use the same version of the VIC model. Secondly, in a preliminary assessment of snow water equivalent (SWE) simulation skills at select SNOTEL sites in the WUS, we found that VIC 4.1.2 demonstrated superior performance compared to VIC5. This finding, coupled with our research group's extensive experience and proven results with VIC 4.1.2, informed our decision to use this version. For WRF-HYDRO, we utilized the most current version to benefit from the latest advancements in the model. We have clarified these aspects in our revised manuscript, ensuring a comprehensive understanding of our model version selection rationale.

We added sentences explaining this at lines 203-210 and lines 223-224 in the revised manuscript.

-Figures 9-10-11 can be given in appendix.

Thank you, we moved them to the appendix in the revised manuscript.

-The paper needs a separate Discussion section and a separate Conclusions (bullets) section. Summary can be appropriate for “engineering corps” reports not for HESS papers.

Thank you for your suggestion; accordingly we have restructured our manuscript to include a distinct discussion section (lines 555-596). This section examines the significance of our findings and their potential limitations, and also suggests directions for future research. We have separated the conclusions (lines 597-618) into a separate section which summarizes our main findings, contributions of our work, and its practical implications.

References:

Bass, B., Rahimi, S., Goldenson, N., Hall, A., Norris, J. and Lebow, Z.J.: Achieving Realistic Runoff in the Western United States with a Land Surface Model Forced by Dynamically Downscaled Meteorology. Journal of Hydrometeorology, 24(2), 269-283, 2023.

Gochis, D. and Coauthors: Overview of National Water Model Calibration: General strategy and optimization. National Center for Atmospheric Research, accessed 1 January 2023, 30 pp., https://ral.ucar.edu/sites/default/files/public/9_RafieeiNasab_CalibOverview_CUAHSI_Fall019_0.pdf, 2019.

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.

Livneh B, Rosenberg, E.A. , Lin, C. , Nijssen, B. , Mishra, V. , Andreadis, K. , Maurer, E.P. and Lettenmaier, D.P.: A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States: Updates and extensions, Journal of Climate, doi:10.1175/JCLI-D-12-00508.1, 2013.

Naeini MR, Analui B, Gupta HV, Duan Q, Sorooshian S. Three decades of the Shuffled Complex Evolution (SCE-UA) optimization algorithm: Review and applications. Scientia Iranica. 2019;26(4):2015-31.

3. Response to reviewer 2

General comments

This study presents the calibration and regionalisation of two land-surface hydrologic models in 263 catchments in the Western US. The results indicate that the median Kling-Gupta efficiency obtained in model calibration and regionalisation outperforms earlier/baseline study.

The study focuses on an interesting topic, but in its current form, it reads more like a technical report than a research paper. The Introduction nicely presents the context (i.e. why it is important to simulate water cycle components in the study region accurately). Still, the synthesis and formulation of the current research gaps need to be significantly improved. There is a large body of literature focusing on regional calibration of hydrologic models, transfer of model parameters at the regional scale, definition of the signatures used for similarity definition etc. It needs to be made clear how this study goes beyond the existing studies. The formulation of the research questions needs to be very precise and linked with presenting the research gaps. In its current form, it needs to be clarified whether the main goal is to propose and evaluate some methodological advance in model calibration and/or regionalisation or to present some new factual information about the study region (a case study analysis).

Thank you for your insightful comments. Streamflow forecasts are crucial for promoting sustainable water practices and building resilience to water-related challenges, and robust hydrological model simulations are at the core of streamflow forecasts. The calibration of parameters, although time-consuming and computationally expensive, is key to enhancing model performance. Our study

acknowledges a gap in the availability of finely-tuned, high-resolution calibrated parameter sets for the Noah-MP and VIC models in the Western United States (WUS).

To address this gap, our study applies globally optimized calibration across 263 river basins in the WUS, at a fine resolution of 1/16 degree latitude-longitude. Utilizing the VIC and Noah-MP models, our approach extends beyond the existing scope of hydrological studies. We further regionalized these calibrated parameters to all 4816 HUC-10 basins in the WUS, developing high-resolution parameter sets. These sets are intended to bolster regional hydrological studies and climate change assessments, offering significant benefits for water resource management and environmental planning.

In response to your feedback, we enhanced our manuscript by elaborating on how our study fills an existing gap in globally calibrated hydrological model parameters at a fine spatial resolution and on the extensive spatial scope, which includes calibration at the above-mentioned 263 river basins in the WUS and the regionalization across 4816 WUS HUC-10 basins. Furthermore, our application of two widely used hydrological models, VIC and Noah-MP, introduces an additional layer of complexity and relevance. The use of these models allows us to address a broader range of uncertainties associated with hydrological modeling in varied climatic and geographic contexts. Accordingly, we have made revisions to the introduction section at lines 77-108 in the revised manuscript.

The selection of the two models needs to be better justified. What are the differences in runoff generation between the models (and how is it linked with the regional variability of runoff generation in the study region)? It needs to be clarified why to use a 3hr simulation time step, when model inputs are daily. It is also not clear why to calibrate only selected soil-related parameters and how the selection is linked

with the runoff generation processes and their variability in the study region. For example, are the snow accumulation and melt processes less important? Or are the snow-related model parameters already accurately calibrated? More importantly, the results and the differences between the two models need to be better linked with the main runoff generation processes (and their regional variability).

I missed the discussion of the results, which will link the new findings with previous studies. This can enhance the demonstration of the novel scientific contribution of the study.

Thank you, we answer your comments below:

(a) Why we selected these two models.

We chose to focus on the Variable Infiltration Capacity (VIC) model and the Noah-Multiparameterization (Noah-MP) LSM due to widespread previous application of these two models both in the U.S. and globally, as highlighted by Mendoza et al. (2015) and Tangdamrongsub (2023).

Our rationale for incorporating two distinct hydrological models lies in addressing the inherent variability and uncertainty in such simulations. By using two models, we aim to enhance the robustness of our study and better encompass structural uncertainties.

The VIC model is renowned for its widespread popularity (the original reference, Liang et al., JGR 1994 has been cited almost 3000 times) and demonstrated success in simulating runoff on a global scale (e.g. Adam et al 2003 & 2006; Livneh et al 2013; Schaperow et al 2021). Its established track record makes it an invaluable component of our analysis. The Noah-MP model is relatively newer, but is the hydrologic core of the National Water Model (NWM) which is being used increasingly domestically and internationally.

Further reinforcing our choice is a study by Cai et al. (2014), which evaluated the hydrologic performance of four LSMs in the contiguous United States using the North American Land Data Assimilation System (NLDAS) test bed. This study found that Noah-MP exhibited superior performance in soil moisture simulation and ranked highly in Total Water Storage (TWS) simulations. Conversely, the VIC model was distinguished for its excellence in streamflow simulations.

Our decision to utilize both the Noah-MP and VIC models is predicated on their proven effectiveness in simulating a wide range of hydrological processes. The unique runoff generation methodologies of each model are particularly pertinent for capturing the diverse hydrological characteristics of the WUS. This methodological diversity allows us to more comprehensively assess runoff generation mechanisms and their spatial variability within the region. We revised the manuscript in section 2.2 Land Surface Models at lines 151-224 to address more on why we selected these two models.

(b) Differences in runoff generation between the models.

Noah-MP has four runoff physics options and after evaluation we decided that the free drainage exhibited the most substantial performance enhancement after calibration. As a result, we chose to continue using this option which is incorporated in the NWM. This runoff physics option is signified with infiltration-excess based surface runoff scheme and gravitational free-drainage subsurface runoff scheme [Schaake et al., 1996]. Noah-MP has four soil layers and each layer has a fixed depth (from top to bottom, 0.1m, 0.3m,0.6m,1.0m).

In VIC, each grid has up to three soil layers and the depth can be different for each grid cell. The infiltration into the top-most layers is controlled by variable infiltration capacity (VIC) parameterization (Liang et al., 1994). The flow is gravity-driven from upper layers to lower layers (Brooks and Corey, 1964). The function of

the soil moisture in the third layer is linear below a soil moisture threshold and becomes nonlinear above that threshold. [Liang et al., 1994]. (could be seen as a combination of infiltration excess and saturation overland flow combination Liang and Xie (2001)). We revised the manuscript to include these differences in the runoff generation between the models at lines 164-174.

(c) How is it linked with the regional variability of runoff generation in the study region?

Both Noah-MP and VIC show good baseline performance along the Pacific Coast, in central to northern CA. Those areas have a high runoff ratios (specifically spring and annual runoff ratio) and high mean winter precipitation and mean annual max daily precipitation. These features suggest runoff physics that are dominated by the saturation excess mechanism, thus both VIC and Noah-MP perform well in these regions. VIC's baseline KGE generally is high in the inland northwest which has lower mean annual max daily precipitation and deeper groundwater table, VIC might be better to simulate these basins because it has varied soil moisture depth while Noah-MP has fixed soil moisture depth. Substantial post-calibration improvements occurred for both models in most areas, especially in regions where the baseline KGE was low, such as southern CA and the southeastern part of the study region. We revised our manuscript to include these points at lines 388-397.

(d) Why use a 3hr simulation time step, when model inputs are daily?

The choice of a 3-hour simulation time step, despite having daily model inputs, was intended to capture the diurnal cycle of energy balance and hydrological processes, which can be significant in regions with large variations in daily temperature and solar radiation. This finer temporal resolution aids in better representing the hydrological response and energy dynamics, especially in snow-dominated catchments. Besides that, we did an analysis on the timestep of Noah-MP

and found that at least of 3-hour timestep is needed to generate robust simulations. Although VIC can be run at a daily time step, Noah-MP generally is implemented with a sub-daily times step, so to make the two models comparable, we run VIC simulations at the same 3-hour time step as Noah-MP. We've made revisions in the model section at lines 244-249 to address these points.

(e) why calibrate only selected soil-related parameters and how the selection is linked with the runoff generation processes and their variability in the study region?

Our focus on calibrating soil-related parameters was based on their critical role in runoff generation. We aimed to address the key processes such as infiltration, soil moisture storage, and (in the case of Noah-MP) groundwater recharge, which are pivotal in the WUS's diverse hydroclimatic settings. We prioritized the calibration of these parameters to improve the representation of soil-water interactions, a major driver of runoff variability in the region. Concerning snow accumulation and melting processes, we acknowledge their importance. Before we conducted the calibration, we conducted snow simulation verification at 20 selected Snotels sites across WUS. Our assessment indicated that the existing parameterizations for snow processes in both models were adequate for our study region. We've revised our manuscript in the calibration section at lines 258-267 to clarify these points.

(f) Discussion of the results.

Contextualizing Findings with Previous Research: To demonstrate the novel scientific contribution of our study, we expand our introduction at lines 94-99. We added a separate discussion section at lines 555-596. We appreciate the opportunity to refine our manuscript based on your feedback, and we believe these revisions will significantly enhance the clarity and impact of our research.

References:

Adam, J.C. and Lettenmaier, D.P.: Adjustment of global gridded precipitation for systematic bias, J. Geophys. Res., 108(D9), 1-14, doi:10.1029/2002JD002499, 2003.

Adam, J.C., Clark, E.A. , Lettenmaier, D.P. and Wood, E.F.: Correction of Global Precipitation Products for Orographic Effects, J. Clim., 19(1), 15-38, doi:10.1175/JCLI3604.1, 2006.

Brooks, R.H. and Corey, A.T. (1964) Hydraulic Properties of Porous Media. Hydrology Paper, Vol. 3, Colorado State University, Fort Collins.

Cai, X., Yang, Z.L., Xia, Y., Huang, M., Wei, H., Leung, L.R. and Ek, M.B., 2014. Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed. Journal of Geophysical Research: Atmospheres, 119(24), pp.13-751.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A Simple hydrologically Based Model of Land Surface Water and Energy Fluxes for GSMs, J. Geophys. Res., 99(D7), 14,415-14,428.

Liang, X. and Xie, Z., 2001. A new surface runoff parameterization with subgrid-scale soil heterogeneity for land surface models. Advances in Water Resources, 24(9-10), pp.1173-1193.

Livneh B, Rosenberg, E.A. , Lin, C. , Nijssen, B. , Mishra, V. , Andreadis, K., Maurer, E.P. and Lettenmaier, D.P.: A long-term hydrologically based data set of land surface fluxes and states for the conterminous United States: Updates and extensions, Journal of Climate, doi:10.1175/JCLI-D-12-00508.1, 2013.

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.

NOAA (National Oceanic and Atmospheric Administration): National Water Model: Improving NOAA's Water Prediction Services, 2016.

Schaperow J.R, Li, D., Margulis, S.A., Lettenmaier D.P. :A near-global, high resolution land surface parameter dataset for the variable infiltration capacity model. Scientific Data. Aug 11;8(1):216, 2021.

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.

4. Response to student reviewer

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. 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 have enhanced 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 at lines 150-190; (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 at lines 87-90,164-174 and 384-397.

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.

We've added a clearer explanation of the parameter calibration and regionalization processes in the revised MS in section 3&4 at lines 253-355 and lines 415-514. 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. We added comments to this effect at lines 589-596.

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.

To address your comment, we added a more detailed explanation of this process in the revised MS at lines 430-506. 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 9 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 has 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 examined the effects of alternatives (to KGE) objective functions at lines 349-355 and 579-584.

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.

We apologize for any missing citations in the reference section, which we have corrected.

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. We now comment on this effect at lines 589-596 and 598-602.

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

Thank you, we've improved this.

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

Done.

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

Thank you, we’ve fixed it.

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

We’ve added 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).

Done.

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).

The best feature combinations and numbers differ for different models and in different regions. We explored this 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 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 at lines 416-507.

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.”

Edited as you suggest.

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

We’ve added more discussion on the limitations of our study. This includes selection of calibration objective function and metrics, the potential limitation of regionalization method, the applicability of our findings to other hydrological contexts. 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.