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. We recognize the importance of clarity in our experimental framework, especially regarding our choice of different algorithms for the VIC and Noah-MP models.

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.

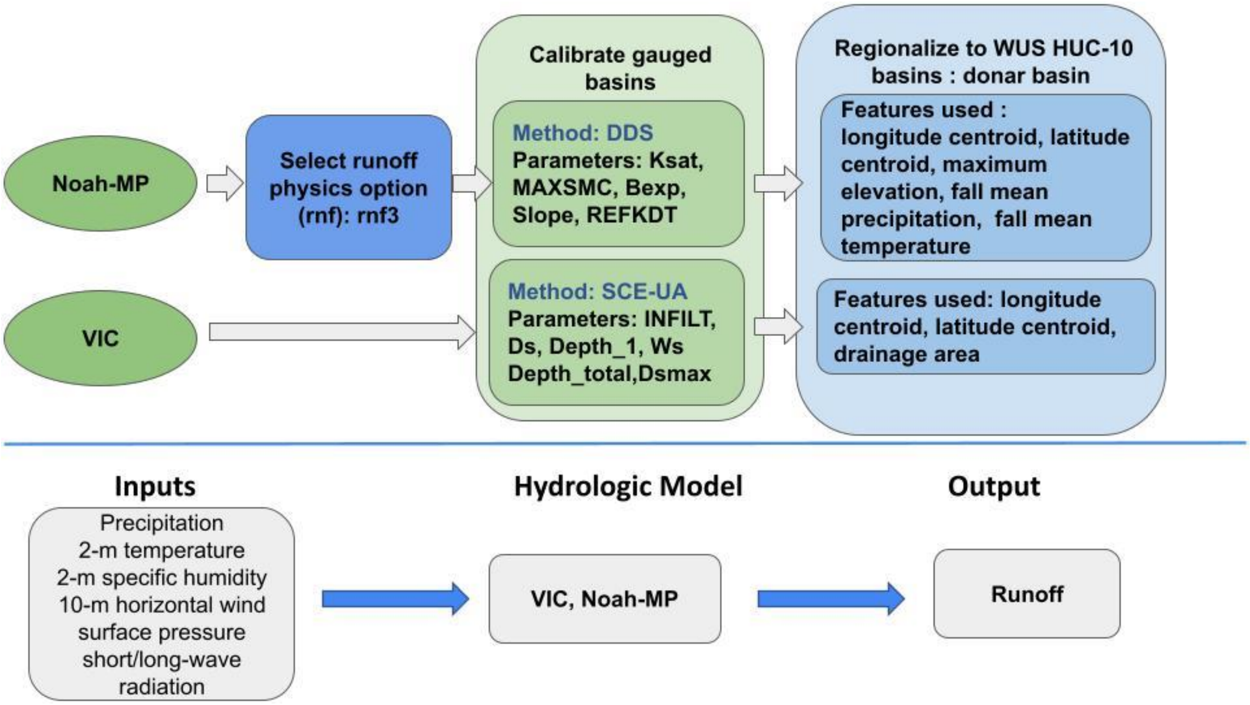
On the other hand, the Noah-MP model is more computationally demanding, and required a different approach. For this reason, we selected the Dynamically Dimensioned Search (DDS) algorithm. DDS is also used in the CONUS implementation of the National Water Model, which uses Noah-MP as its hydrologic core (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 it 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 will include a more detailed explanation (similar to the above) in our revised manuscript to better explain our methodological choices and their rationale.

Please a conceptual diagram/framework describing your methodology.

To improve clarity and understanding of our research design, we will include a diagram in our revised manuscript similar to the one below. This diagram illustrates our calibration framework, delineating the distinct but complementary approaches we took for each model.


Noah-MP
VIC
Select runoff physics option (rnf): rnf3
Calibrate gauged basins
Method: DDS
Parameters: Ksat, MAXSMC, Bexp, Slope, REFKDT
Method: SCE-UA
Parameters: INFILT, Ds, Depth_1, Ws Depth_total,Dsmax
Regionalize to WUS HUC-10 basins : donar basin
Features used : longitude centroid, latitude centroid, maximum elevation, fall mean precipitation, fall mean temperature
Features used: longitude centroid, latitude centroid, drainage area
Inputs
Precipitation
2-m temperature
2-m specific humidity
10-m horizontal wind
surface pressure
short/long-wave radiation
Hydrologic Model
VIC, Noah-MP
Output
Runoff

-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 acknowledge the importance of using diverse metrics to comprehensively assess the performance of the VIC and Noah-MP models. In this research, 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 intend to include (in the appendix) maps of (normalized values of) multiple performance measures, which will allow direct comparisons of other measures with KGE.

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. In the revised paper, we will expand slightly our discussion of high vs low flow performance.

-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 will incorporate additional sentences about this analysis in Section 3.1 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.

-Figure4: 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 remote sensing products from MODIS, SMAP, SMOS, ESA, and ALEXI offer valuable data 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 such diverse 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 will revise the manuscript to comment on this broader perspective and acknowledge the importance of integrating remote sensing data in hydrological modeling.

-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

In this study, 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 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.

-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 will clarify these aspects in our revised manuscript, ensuring a comprehensive understanding of our model version selection rationale.

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

Thank you, we'll move 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 suggestions. We agree that a clear distinction between the Discussion and Conclusion sections would enhance the readability and academic rigor of the paper. In line with your advice, we will restructure our manuscript to include a distinct discussion section. This section will be dedicated to examining the significance of our findings, their relevance to the broader hydrological research community, and their potential limitations. It will also suggest directions for future research and provide a deeper interpretation of our results. We will also develop a separate conclusions section, designed to be brief yet informative. This section will succinctly summarize the main findings, contributions of our work, and its practical implications. It will be crafted to ensure that the central messages of our study are easily understood and remembered by our readers. We appreciate your guidance on this matter and are committed to making the necessary adjustments to improve our manuscript.

References:

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.