Responses to the Comments of Reviewer #1 on ⟨hess-2022-282⟩

Peishi Jiang Pin Shuai Alexander Sun Maruti K. Mudunuru
Xingyuan Chen

December 28, 2022

> General comments:
> This study showcases a deep learning optimization method for a high-resolution hydrologic model supported by information theory. I appreciate the honest evaluation of the methodology, in-depth reasoning of the deteriorating model performance for ET, and examination of results and conclusions aligned with earlier studies. In general, this paper is well-written with a novel contribution. However, I think the paper would be stronger if the authors can address the following comments.

Thank you for the thorough review of our manuscript. We addressed each comment as shown below.

> Model validation for climate sensitivity: Currently, the model validation period overlaps with the period for calibrating ATS parameters. I am curious whether the optimized parameters would be able to capture the climate sensitivity on flow and ET, i.e., improving the flow/ET performance outside of the calibrating period (2016-2019). It would strongly support this tool's eligibility in climate change studies.

We now extend the simulation period to 31-12-2021, which is the end of the available Daymet forcing (see Figure 1(b)). We split the whole period into the calibration (1-10-2016 through 30-9-2019) and the evaluation (1-10-2019 through 31-12-2021) periods, used to calibrate and evaluate the climate sensitivity of the model, separately. The result in Figures 9 and 10 shows that the performances of the calibrated ATS during the evaluation period are very close to that of the calibration period, suggesting the adaptability of the estimated parameters to an uncalibrated time period. We revised the associated results and discussion as follows:

" (L232-L239) Here, we separate the entire observations in Figure 1(b) into model calibration and evaluation periods in order to assess the adaptability of the estimated parameters to an uncalibrated period. To this end, we calibrate ATS only using the simulations during water year 2017 to water year 2019 and used the remaining observations (till 31 December 2021) for model evaluation. The ensemble runs used for sensitivity analysis and inverse modeling are performed during the calibration period. The calibrated ATS forward runs were then performed on both periods and compared against the observations in Figure 1(b). We assess the performances of the calibrated models on both periods by using two scale-independent metrics: the Nash-Sutcliffe Efficiency (NSE; [4]) and the modified Kling-Gupta Efficiency (mKGE; [3]).

...

(L346-L352) **Adaptability of the calibrated model in the evaluation period.** For both Q and ET, both NSE and mKGE of the evaluation period (the cyan lines) are astonishingly close, if not identical, to that of the calibration period (the blue lines). Whenever the calibrated model shows improvement using the knowledge-informed inverse mapping (such as the comparison between qonly-3yrs and mi-qonly-3yrs), we can observe the corresponding improvement in the evaluation period. Such consistent performance between the two periods suggests the robustness of the estimated parameters to climate sensitivity. "

> ET from flux tower: In this study, the authors have demonstrated that worse ET performance results from poor quality of MODIS ET products. In this study region, is there ET data from the flux tower that could be used for implementing this workflow? Even though the flux tower ET data has less spatial coverage, the data quality can be better, which might be more useful than MODIS ET when calibrating hydrologic parameters.

We looked into the AmeriFlux and there is no flux tower site available in this watershed. Therefore, we are not able to perform the calibration against the site-based observations.

> Specific comments:
> L158: Can the authors elaborate on what five soil types and four geological types are?

They are grouped subsurface characteristics in the soil and geological layers using k-means clustering. We add the following description for a better elaboration (L161-L163): "Each clustered soil or geological type is associated with a specific set of subsurface characteristics (such as permeability), which are assigned to the corresponding grouped grid cells. These subsurface characteristics are important in controlling flow dynamics and can be estimated from hydrological observations."

> L160: A 1000-year spin-up is extremely long. Can the authors briefly explain the reason for this long spin-up even if it might be explained in Shuai et al 2022?

We have revised the associated text to clearly explain the motivation for the 1000-year cold or steady-state spinup (note that the cold spin-up took less than one hour to complete on 128 CPU cores due to the faster model convergence once it reached quasi-steady-state.): " (L163-L167) To ensure that the model achieved a physically appropriate initial state, two spinups were performed sequentially, including (1) a cold spinup that ran the model for 1000 years by using constant rainfall and led to steady-state condition at the end of the simulation (e.g., converged total amount of subsurface water storage) and (2) a warm spinup that was initialized by the steady-state spinup result and performed a transient simulation for 10 years (i.e., 1 October 2004 – 1 October 2014) under the Daymet forcing. "

> L162: Could the authors briefly explain how they preselected the parameters in this study?
> L208: Does the MI have to be zero? If the MI between a parameter and the model responses is small enough, is it possible to neglect that parameter? What would be a proper threshold for it?
> L249-250: Given the narrowed list, it seems that the authors eliminated the parameters with small MI (not zero), which slightly contradicts the previous statement where only parameters with zero MI would be eliminated (L208). It would be helpful to clarify the threshold of MI below which the parameters will be eliminated.

As all three comments are associated with how we performed the preliminary sensitivity analysis using mutual information, we reply to them in one thread here. In short, the preselection is based on the mutual information (MI) computed for each parameter and each response at a given time step. For a given model response (e.g., Q), we say it is sensitive to a parameter if the proportion of non-zero MI over all the time steps is greater than a given threshold (i.e., 5% in this study). For each MI calculation, we performed a statistical significance test to determine whether the computed MI is significant and set MI to zero if the test fails. So, the MI can be zero. We enriched the description in the associated texts as below:

“

“ (L190-L194) In this study, we follow a similar strategy of [1] to estimate p using 10 evenly divided bins along each dimension and perform SST tests to filter out any non-significant MI value with a significance level of 95% based on 100 bootstrap samples. In other words, the computed MI is set to zero if the statistical significance test fails.

...

(L205-L209) This preliminary MI analysis would allow filtering out the parameters that show little sensitivity to the model responses, thus reducing the number of parameters to be calibrated. This filtering process is performed based on whether a parameter demonstrates sufficient sensitivity across the simulation period. In this study, we selected the parameter whose proportion of the non-zero MI is larger than 5% of the overall time steps for the following full sensitivity analysis.

...

(L285-L290) Based on the proportion of nonzero MI over all the time steps (see Figure A1 in the appendix), we find that Q is mostly sensitive to (using a threshold of 5%) perm_s3, perm_s4, perm_g1, perm_g4, snowmelt_rate, snowmelt_degree_diff, and priestley_taylor_alpha_transpiration, and ET is mostly sensitive to priestley_taylor_alpha_transpiration, priestley_taylor_alpha_snow, perm_s3, perm_g1, and perm_g4. Consequently, we narrow down the parameters to be calibrated by taking the union of the two sets of parameters that show sensitivities to either Q or ET (also highlighted in Table A1). ”

> `L208-210: Interesting! Great summary!`

Thank you for the generous comment.

> `L215: When training using different combinations of years, why do the authors only look at Q, not ET?`

We do not use ET for multi-year analysis because the extrapolation issue of the ET observations deteriorates the parameter estimations using the inverse mapping, as described in Section 3.3. In other words, a multi-year analysis including ET would be questionable and not trustworthy to evaluate the impact of dry and wet years. Therefore, we performed the multi-year analysis against only Q.

> `L286-287: Please clarify whether the extrapolation issue partially or solely contributes to the worse MI-informed results.`

The inferior calibrated ATS runs using knowledge-informed deep learning are attributed to both the extrapolation issue of the observations and the potential high uncertainty of the ET product. The associated texts are described below:

“ (L354-L366) This surprising result is probably attributed to both the extrapolation issue of ET observations and the high uncertainty of the remote sensing product. Compared with the ensemble simulation of Q (Figure 5(a)) that captures most observed Q, a majority of ET observations exceed the range of the ATS ensemble of ET during the low ET period each year (i.e., wet seasons or September through May next year; see Figure 5(b)). While it is possible that the defined sampling ranges of the two Priestley Taylor coefficients in Table A1 are too limited to provide sufficient variations of ET dynamics, the uncertainty of the MODIS ET product also plays a role here [2, 6]. [6] show that the MODIS ET product has much poorer performance and higher uncertainty in the Colorado Basin than in most of the remaining areas in the United States. The large uncertainty of this remote sensing product probably results from the increasing error in the satellite data caused by the cloudier sky in the mountainous region [5], particularly during the dry seasons (i.e., May through September) [6]. In other words, although the ET ensemble gives a better coverage on the observations in the dry seasons than the wet seasons (Figure 5(b)), that could be due to the underestimation of the MODIS ET in the dry period with high ET such that the mismatch between the ET ensemble and the observed ET could be probably more significant. ”

> `Author name: Should the third author be Alexander?`

Dr. Sun’s first name is corrected now.

References

[1] P. Jiang, K. Son, M. K. Mudunuru, and X. Chen. Using mutual information for global sensitivity analysis on watershed modeling. *Water Resources Research*, 58(10), 2022.

[2] M. S. Khan, U. W. Liaqat, J. Baik, and M. Choi. Stand-alone uncertainty characterization of gleam, gldas and mod16 evapotranspiration products using an extended triple collocation approach. *Agricultural and Forest Meteorology*, 252:256–268, 2018.

[3] H. Kling, M. Fuchs, and M. Paulin. Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424-425:264–277, 2012.

[4] J. Nash and J. Sutcliffe. River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.

[5] G. B. Senay, S. Bohms, R. K. Singh, P. H. Gowda, N. M. Velpuri, H. Alemu, and J. P. Verdin. Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the sseb approach. *JAWRA Journal of the American Water Resources Association*, 49(3):577–591, 2013.

[6] T. Xu, Z. Guo, Y. Xia, V. G. Ferreira, S. Liu, K. Wang, Y. Yao, X. Zhang, and C. Zhao. Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous united states. *Journal of Hydrology*, 578:124105, 2019.