Responses to the Comments of Reviewer #2 on ⟨hess-2022-282⟩

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

December 28, 2022

> This study aims at basin scale parameter calibration for a physical hydrologic model (ATS) using DL-based inverse method. The authors leveraged the mutual information (MI) for the global sensitivity analysis to identify the relation between parameters and model simulations, which was later applied to the input selection of a MLP parameter inverse model. They executed different groups of simulations and analyses to comprehensively evaluate the proposed framework. The MS is well-written with overall structure easy to follow. I provide my suggestions below regarding better clarifying several points and hopefully they can be useful to further improving the quality of this study.

Thank you for the accurate summary and we appreciate your careful reviews and comments.

> As my understanding on this study, the title ‘‘knowledge-informed DL’’ is mainly represented by the MI sensitivity analysis used in the input selection for the following inverse modeling. Knowledge informed learning, generally in my mind, is applying physical laws or constraints to the data driven model based on our domain knowledge. To bridge the proposed MI and physical processes together and better strengthen the headline of this study, I suggest the authors try to link the MI results with physical processes of the study area and give some physical explanations of the results from sensitivity analysis. This can further highlight the physical representations of this study.

We have added the following description to better delineate the knowledge obtained from the sensitivity analysis which further facilitated the follow up inverse mapping development:

“ (L304-L314) **Physical knowledge obtained by MI analysis.** The sensitivity analysis reveals the seasonal importance of these watershed characteristics to the hydrological fluxes in this area (Figure 5). During the low flow period (September through March of next year), Q is mostly controlled by the subsurface permeability (i.e., perm_g1, perm_s3, and perm_s4) which regulates both the infiltration and the groundwater movement. Transpiration also plays a role in driving the low flow dynamics through the Priestley Taylor coefficient (e.g., priestley_taylor_alpha_transpiration). During the high flow period (March through September), the snow melting process turns out to be the most critical factor in contributing to the large runoffs, which complies with the prior knowledge about the dominance of the snow process in this watershed. Likewise, the total ET is by and large attributed to a variety of evaporation and transpiration. Snow evaporation is the main component of the total ET in both late autumn and winter when the snow melting rarely happens. On the other hand, in warmer and high-flow seasons, transpiration becomes the dominant contributor to the total ET. The seasonable pattern of the sensitivity of each parameter not only uncovers the hydrological process in the watershed but also serves as the basis to select the most informative model responses to estimate each model parameter. ”

> I am still confused at the details about how the inverse framework is set up and trained. My understanding is that you first run some simulations with ATS (how are the parameters first initialized here?) and use the simulations and parameters to train an inverse mapping with inputs selected by MI, and then replace ATS simulations with real observations to estimate parameters. Does the \responses" mentioned throughout the paper mean the simulated ATS discharge and ET? What are the training targets and how do you develop the structure, tune the hyperparameters and train the DL framework? What are the training and testing dataset separation?... Maybe I didn't understand some parts very well, but indeed expect the authors can better clarify their methodology and results to make readers more easily understand this work.

Correct, we generated ensemble simulations of ATS to perform both MI-based global sensitivity analysis and develop the deep learning (DL)-based inverse mappings. The mappings, developed using multilayer perceptrons (MLPs), estimate model parameters from model responses that refer to streamflow and ET. The technical details of DL model development were described in the appendix of the preprint version. For better readability, we now moved it to Section 2.4 of the main manuscript and revised the associated texts as follows:

" (L240-L265) For comparison purposes, we developed both the original inverse mapping and our proposed knowledge-informed version for parameter estimation. While a separate neural network is developed for estimating each parameter by using knowledge-informed inverse mapping (Figure 3(b)), the original inverse mapping estimates all parameters using one neural network and is developed by following the same strategy in [1] and [3] (Figure 3(a)). Further, to assess the impact of different responses in calibration, we developed three types of inverse mappings that take various model responses: (1) using both Q and ET; (2) using only Q; and (3) using only ET. Additionally, a multi-year analysis was performed by training inverse mappings using Q of different combinations of observed years to evaluate both the impacts of the dry versus wet years and the number of observed years used in calibration.

All the inverse mappings developed in this study are listed in Table 1. Each mapping was developed using a multilayer perceptron (MLP) model as follows. The input of an MLP is an array concatenating the responses to be assimilated within a given calibration period. The output is the model parameter(s). Let's denote the number of input neurons, output neurons, and hidden layers as N_i, N_o, and N_l, respectively. N_i depends on the type of inverse mapping (with or without being knowledge guided), the selections of the response variable(s), and the number of calibration years, varying from ∼100 using one year of Q to 1,785 using all three years of Q and ET. N_o equals either one (i.e., estimating each parameter using knowledge-informed DL calibration) or the number of all the parameters (i.e., using inverse mapping without mutual information). Given N_i, N_o, and N_l, we adopt the arithmetic sequence to determine the number of neurons at each hidden layer $N_{h,l} = \lfloor N_i - \frac{N_i - N_o}{N_l} \times l \rfloor$ (where $1 \leq l \leq N_l$ and $\lfloor \bullet \rfloor$ is the floor function). In doing so, the information from a sequence of observed responses can be gradually propagated to estimate the parameters. We use the leaky ReLu as the nonlinear activation at the end of each layer. Based on the order of the Sobol sequences, we sequentially split the 396 realizations into 300/50/46 for train/validation/test sets, respectively, such that each set is able to cover the full range of the parameter ensemble as much as possible. We trained each MLP using mean square error (MSE) as the loss function over 1,000 epochs with a batch size of 32. The Adam optimization algorithm, a stochastic gradient descent approach, was used to train the neural network. We performed hyperparameter tuning on each MLP using grid search to find the optimal result by varying the number of hidden layers $N_l = [1, 3, 5, 7, 9, 10]$ and the learning rate $l_r = [1e-5, 1e-4, 1e-3]$. The performances of these mappings are further evaluated on the two magnitude-independent metrics, NSE and mKGE. To have consistent comparisons between mappings with and without being knowledge guided, both metrics are computed for the estimation of each parameter based on the test dataset. "

> I didn't understand the result of Figure 7 well and hope the authors can give more explanations. Which variables are the NSE and mKGE calculated on, estimated parameters or model simulations? If they are simulation metric, are theses simulations from the model forwarding with parameters estimated from real observations (Q & MODIS ET inverse)? For each individual parameter evaluation, how do you set up the values of other parameters when doing ATS forwarding. The caption notifies the performance is reported on testing data, but I didn't see how the authors divide testing and training data.

The result of Figure 7 (now Figure 8) was calculated between the true parameters and the estimation by inverse mappings in the test dataset. How the train/validation/test data were splitted is described in the reply to the previous comment regarding the details of DL model development. We revised the caption of Figure 8 as: "Parameter estimation performance of the developed deep learning (DL) inverse mappings on the test dataset using the model responses in the calibration period with regards to (a) the modified Kling-Gupta Efficiency (mKGE) and (b) the Nash-Sutcliffe Efficiency (NSE). Green and light blue represent the mappings without and with being knowledge informed, respectively. Blank, cross, and circle textures are used to represent the mapping using discharge only (qonly), evapotranspiration only (etonly), and both (qet), respectively.".

> I am thinking this multiple-years training VS one-year training discussed in section 3.3. As for multiple years, you choose to increase the input neuron number, or keep the one-year structure not changed and just use multiple years data as more training samples? I think the latter one could be more beneficial because inputting three-year time series once to the model would require large amounts of parameters in the input layer which can be inefficient and overfitted to small training data.

For multiple years, we increased the number of inputs of the DL model and performed hyperparameter tuning to find the optimal architecture of the model. The tuning result partially addressed the overfitting issue. Indeed, we observe a limited impact of overfitting from the training result (see Figures 7, A3, and A4). Therefore, we did not try a different model architecture which complicates the hyperparameter tuning procedure and is out of the scope of this study. We have added the following in the result section to demonstrate this point:

" (L316-325) The developed inverse mappings demonstrate limited overfitting issues. Figure 7 plots the training and validation loss over epochs of the seven parameters, each of which is estimated by the knowledge-informed inverse mapping using the corresponding three years of sensitive streamflows (i.e., mi-qonly-3yrs). It can be observed from the figure that both losses quickly decrease with epochs with little discrepancies. Particularly, the parameters sharing with higher mutual information with streamflows show faster convergences of the loss function and do not have overfitting problem (e.g., perm_s3 and snowmelt_degree_diff; see Figure 6(a)). The discrepancy between training and validation losses gets slightly larger for less sensitive parameters (e.g., perm_g4) where streamflow is less informative in parameter estimation. Indeed, informative model responses can provide better parameter estimations, thus reducing the overfitting impact. The limited impact of overfitting is also evident from the NSE and mKGE barplots of the training, validation, and test sets of all the inverse mappings (see Figures A3 and A4), where most mappings have similar performances on parameter estimations among the three sets. "

> Another point I would be interested in is whether the authors have tried adding meteorological forcings to the inputs of inverse modeling. I feel the forcing-hydrologic response pair is very important to inform the characteristics of basin processes reflected in model parameters. I am expecting the paired input may bring more benefits to this study.

Including atmospheric forcing in the inputs of the neural network might be inappropriate for this basin-specific study. This is because the forcings do not change with ensemble realizations and thus are constant values to the DL model inputs, which might even deteriorate the DL model performance. Including such basin characteristics would be more beneficial to studies encompassing multiple basins. We thus added it as one future work in the conclusion:

" (L458-L460) One potential future work is to develop a unified inverse modeling framework for multiple basins, where the atmospheric forcings and basin characteristics can be also used as the inputs of the inverse mappings in addition to the realization-dependent model responses. "

> Line 76 Do you intend to discuss the overfitting problem here? Large number of weights and limited realizations as training data may cause overfitting with a complicated model.

We now discuss the overfitting problem as follows:

" (L77-L79) Further, when using all observed responses as inputs, the potentially large amount of trainable weights of the DL model can make the model training hard and cause the overfitting of the model [4], thus calling for more realizations used in training. "

> Line 177 Please also give explanations for H(Y|X) to help readers' understanding.

We have added the explanation for $H(Y|X)$ in L184-L185: "$H(Y|X)$ is the conditional entropy that quantifies the uncertainty of Y given the knowledge of X".

> Line 258 and 259 How did the authors safely draw the conclusion of ''improves the MI estimations'' and ''the parameters are falsely considered'' based on the differences of preliminary and full analysis?

This is due to the improved MI estimation of the fully analysis which uses around 400 realizations, as evidenced by the convergence of the MI estimations shown in Figure A2 of the appendix. This converged MI estimation allows us to identify the that is not available in the preliminary analysis. We have revised the associated text to better illustrate this point in:

" (L293-L301) By using more realizations, this complete MI analysis shows a better delineation of parameter sensitivity than the preliminary analysis due to its convergence on MI estimation (see the convergence of the parameter rankings in Figure A2). The convergence on a few hundred realizations is consistent with another MI-based sensitivity analysis study using Soil & Water Assessment Tool (SWAT) [2]. Further, the MI-based parameter ranking suggests that compared with the preliminary analysis, the full analysis (1) improves the MI estimations (e.g., perm_s3); and (2) identifies the insensitive parameters (e.g., perm_s4) that are falsely considered sensitive due to the limited samples in the preliminary analysis (see Figure 6). The main permeability in the soil layer (i.e., perm_s3), for example, now shows higher and more temporally coherent sensitivity to Q (Figure 5(a)). On the other hand, perm_s4, which shows some sensitivity in the preliminary analysis, turns out to be insensitive to both Q and ET with almost zero MI at each time step. "

> Additionally, is it possible that in the preliminary analysis some parameters are not identified but actually behave sensitive if you include them in the full MI analysis?

Yes, it is possible, because the preliminary analysis does not theoretically exclude such false negative cases due to the limited sampling. However, the statistical significance test used to filter the insignificant MI estimation can greatly improve the MI estimation as shown in a previous study in [2], thus partially eliminating such cases. We acknowledge this point in the conclusion:

" (L433-L439) The proposed hierarchical way of sensitivity analysis efficiently utilizes the available limited computational resource through a combination of a prescreening analysis and then a full analysis. Although the prescreening using 50 model runs does not theoretically exclude a false negative case that a sensitive parameter is classified as insensitive, the statistical significance test is able to improve the estimation of mutual information in Figure 4 thus facilitating narrowing down an "accurate" list of parameters to be estimated. Based on the shortened parameter list, a full sensitivity analysis is successfully performed using nearly 400 model runs and provides physically meaningful results on the dependency between the parameters and model responses in Figure 5. "

> Figure 8 The inputs to the inverse model here are real observations or simulated responses?

The forward runs (now shown in Figures 9 and 10) are driven by the parameters estimated by the observations. We have revised the captions of the two figures accordingly.

References

[1] E. Cromwell, P. Shuai, P. Jiang, E. T. Coon, S. L. Painter, J. D. Moulton, Y. Lin, and X. Chen. Estimating watershed subsurface permeability from stream discharge data using deep neural networks. *Frontiers in Earth Science*, 9, 2021.

[2] 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.

[3] M. K. Mudunuru, K. Son, P. Jiang, and X. Chen. Swat watershed model calibration using deep learning, 2021.

[4] X. Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168(2):022022, feb 2019.