Knowledge-Informed Deep Learning for Hydrological Model Calibration: An Application to Coal Creek Watershed in Colorado

Peishi Jiang\(^1\), Pin Shuai\(^1\), Alexandar Sun\(^2\), Maruti K. Mudunuru\(^1\), and Xingyuan Chen\(^1\)

\(^1\)Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory
\(^2\)Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin

Correspondence: Peishi Jiang (Peishi.Jiang@pnnl.gov)

Abstract. Deep learning (DL)-assisted inverse mapping has shown promise in hydrological model calibration by directly estimating parameters from observations. However, the increasing computational demand for running the state-of-the-art hydrological model limits sufficient ensemble runs for its calibration. In this work, we present a novel knowledge-informed deep learning method that can efficiently conduct the calibration using a few hundred realizations. The method involves two steps. First, we determine decisive model parameters from a complete parameter set based on the mutual information (MI) between model responses and each parameter computed by a limited number of realizations (\(\sim 50\)). Second, we perform more ensemble runs (e.g., several hundred) to generate the training sets for the inverse mapping, which selects informative model responses for estimating each parameter using MI-based parameter sensitivity. We applied this new DL-based method to calibrate a process-based integrated hydrological model, the Advanced Terrestrial Simulator (ATS), at Coal Creek Watershed, CO. The calibration is performed against observed stream discharge (Q) and remotely sensed evapotranspiration (ET) from the water year 2016 to 2018. Preliminary MI analysis on 50 realizations resulted in a down-selection of seven out of fourteen ATS model parameters. Then, we performed a complete MI analysis on 396 realizations and constructed the inverse mapping from informative responses to each of the selected parameters using a deep neural network. Compared with calibration using all observations, the new inverse mapping improves parameter estimations, thus enhancing the performance of ATS forward model runs. The Nash-Sutcliffe efficiency (NSE) of streamflow predictions increases from 0.65 to 0.80 when calibrating against Q alone. Using ET observation, on the other hand, does not show much improvement due to both the uncertainty of the remotely sensed product and the insufficient coverage of the model ET ensemble in capturing the observation. By using observed Q only, we further performed a multi-year analysis and show that Q is best simulated (NSE: 0.85) by including in calibration the dry year flow dynamics that shows more sensitivity to subsurface characteristics than the other wet years. Our success highlights the importance of leveraging data-driven knowledge in DL-assisted hydrological model calibration.

1 Introduction

Calibrating a hydrological model is critical to accurately capturing the hydrological dynamics of the simulated watershed, which in turn improves the understanding of the corresponding terrestrial water cycle (Singh and Frevert, 2002). While the in-
creasing complexity and spatio-temporal resolution of the hydrological models enable a better representation of the watershed
dynamics (Kollet and Maxwell, 2006; Coon et al., 2019; Wang and Kumar, 2022), running these models is computationally ex-
pensive (Clark et al., 2017) even with existing high-performance computing resources. This computational burden significantly
impedes the calibration of integrated hydrological models efficiently and accurately.

Balancing the trade-off between computational cost and calibration accuracy is necessary when adopting traditional model
calibration methods (Kavetski et al., 2018). Newton-type optimization methods (Jorge and Stephen, 2006; Qin et al., 2018) are
known for their fast iteration convergence but usually only achieve local optimum. On the other hand, the stochastic methods,
such as the shuffled complex evolution algorithm (Duan et al., 1992), the dynamically dimensioned search algorithm (Tolson
and Shoemaker, 2007), and the ensemble Kalman filter (Reichle et al., 2002; Moradkhani et al., 2005; Evensen, 2009; Sun
and Sun, 2015), are capable of providing better global optimum at the cost of high computational demand. One alternative is
to use a surrogate model that provides fast emulations to replace the physical model during calibration so that one might save
the computational budget while achieving a reasonable calibration result. Mo et al. (2019) employed a dense convolutional
encoder-decoder network as the emulator for a two-dimensional contaminant transport model to estimate the conductivity field
using iterative local updating ensemble smoother. Similar subsurface characterization was also performed by Wang et al. (2021)
who developed a theory-guided neural network as the surrogate of a flow model which was coupled with an iterative ensemble
smoother to estimate the subsurface characteristics. In light of the dimensionality reduction of the model states, Dagon et al.
(2020) calibrated biophysical parameters using a global optimizer on a surrogate that emulates the principle components of
the outputs of community land models. Jiang and Durlofsky (2021) adopt a recurrent encoder-decoder network as the data-
space inversion parameterization to reduce the dimensionality of the model states/parameters and used ensemble smoother
with multiple data assimilation to update the low-dimension latent variables. Despite the successes of using surrogates in
calibration, how to develop an accurate and trustworthy emulator can vary from case to case and, in fact, is still a long-standing
challenge (McGovern et al., 2022).

Recently, Deep Learning (DL)-assisted inverse mapping shows promise in addressing inverse problems and has seen early
successes in hydrology (Cromwell et al., 2021; Mudunuru et al., 2021; Tsai et al., 2021), petroleum engineering (Razak et al.,
2021), and geophysics (Yang and Ma, 2019; Wang et al., 2022). By employing a well-trained DL model (Goodfellow et al.,
2016), this approach maps model parameters from model states/outputs/responses such that once trained, the mapping can
directly infer the parameters based on observations. The inverse mapping outperforms the traditional calibration approaches in
the following ways. First, DL models can better capture the highly nonlinear relationships encoded in the model than ensemble-
based methods, which primarily rely on the linear estimation theory through the Kalman filter (Evensen, 2009; Moradkhani
et al., 2005; Reichle et al., 2002; Sun and Sun, 2015). Yang and Ma (2019) developed a convolutional neural network-based
inverse mapping that outperforms the traditional full waveform inversion adopting the adjoint-state optimization method in
estimating seismic velocity from seismic data. Cromwell et al. (2021) also demonstrate the improved performance of DL-
assisted inverse mapping over ensemble smoother in estimating subsurface permeability used in a watershed model based
on the Advanced Terrestrial Simulator (ATS). Second, training DL models may potentially use fewer realizations than the
traditional methods such as iterative calibration methods that usually require several thousands of realizations to achieve the
model optimization convergence. Mudunuru et al. (2021) show that DL-assisted inverse mapping using 1,000 realizations outperforms dynamically dimensioned search algorithm (Tolson and Shoemaker, 2007) that has to leverage 5,000 realizations in calibrating multivariate parameters of models based on the Soil and Water Assessment Tool (SWAT). Third, the calibration workflow is simpler given that ensemble simulations don’t have to be fully coupled with the inverse mapping. Traditional calibration methods require sophisticated workflows (White et al., 2020; Jiang et al., 2021) to manipulate the model restart (e.g., ensemble Kalman filter), model rerunning (e.g., gradient-based and ensemble-based methods), and the communications between hydrological model and calibration tools, which can be time-consuming. Meanwhile, such an integrated workflow tool is not necessary for developing inverse mapping because model ensemble runs and DL training are now two separate steps. This decoupling of ensemble runs and DL training allows us to use high-performance computing resources to calibrate hydrological models efficiently.

Despite its success, the current DL-assisted inverse mapping is often designed to take all observed states in estimating hydrological model parameters. However, some observational values can be uninformative, or even misinformative, to estimate parameters (Loritz et al., 2018), thus impeding the mapping performance. While the underlying assumption is that the trained DL model can ‘automatically’ delineate the accurate relationship between parameters and observed responses, the limited realizations (e.g., a few hundred) would potentially restrain the DL model from being well trained (Moghaddam et al., 2020). Further, when using all observed responses as inputs, the potential large number of trainable weights of the DL model can make the model training hard and thus calls for more realizations used in training. Lately, several studies proposed new inverse mapping methods that indirectly address this issue by using dimensionality reduction and differential programming. Razak et al. (2021) developed a latent-space inversion that performs the inverse mapping from the model responses to parameters in their reduced spaces through an autoencoder. The dimensionality reduction by using an autoencoder is supposed to not only lower the original high-dimensional data space but also indirectly distill the most relevant dynamics to the parameters. Tsai et al. (2021) leveraged differential programming such that the loss function of an inverse mapping is designed to directly minimize the difference between observations and the responses predicted by a differentiable version of the physical model using the estimated parameters. In doing so, the uninformative responses are automatically given less importance in the loss. Nevertheless, both studies reduce the irrelevant information in implicit ways through complicating the DL-based mapping which, due to their ‘black-box’ nature, does not explicitly show to what extent an observation is relevant to a parameter. Also, by adding another layer of complexity, the inverse mapping can potentially be hard to build. For instance, the current solutions to develop a differentiable physical model rely on either a surrogate or rewriting the model using differentiable parameters (Karniadakis et al., 2021), both of which are research challenges that go beyond addressing the inverse problem itself.

The emergence of knowledge-informed DL provides a new opportunity to resolve the uninformative or misinformative issue by explicitly encoding the complex relationship between the inputs and outputs in the DL model (Willard et al., 2020). Knowledge-informed DL includes, but is not limited to, the following three ways: (1) physics-guided loss function, (2) hybrid modeling, and (3) physics-guided design of architecture. Physics-guided loss function embeds the mathematical relation between inputs and outputs in the loss function, known as physics-informed deep learning (Karniadakis et al., 2021), and has
seen some early successes in earth science. For instance, Jia et al. (2019) leveraged an energy conservation loss in developing a physics-guided recurrent neural network to simulate lake temperature. Hybrid modeling, on the other hand, directly integrates the physical model with the DL model, which often serves as a surrogate for its computationally intense counterpart in the physical model (Kurz, 2021). An example can be coupling a DL-based emulator for turbulent heat fluxes with a process-based hydrological model framework (Bennett and Nijssen, 2021). Lastly, the physics-guided design of architecture explicitly designs the neural networks consistently with prior knowledge. The widely-used convolutional (Atlas et al., 1987) and recurrent (Rumelhart et al., 1986) neural networks fall into this category due to their specific network structures to learn the spatial and temporal relationships, respectively. Other related studies include relating intermediate physical variables to hidden neurons (Daw et al., 2020), explicitly learning nonlinear dynamics through the neural operator (Kovachki et al., 2021), and encoding domain knowledge obtained from nonparametric physics-based kernels into the neural network (Sadoughi and Hu, 2019). Compared with the other two types of knowledge-informed DL, which are usually limited to particular physical dynamics, the physics-guided design of architecture is more generic regarding both the processes of gaining prior knowledge and designing a correspondent neural network.

One important piece of domain knowledge is the pairwise relation between model parameters and responses. That is, how relevant a parameter is to a model response at a given time step. Understanding such a pairwise relationship is essential to select the most relevant model responses to estimate each parameter when building the inverse mapping. To this end, global sensitivity analysis (Razavi and Gupta, 2015; Sarrazin et al., 2016) is a suitable tool due to its capability to quantify the contribution of uncertainty from model inputs and parameters to model outputs, and has been extensively applied in earth system modeling (Hall et al., 2009; Harper et al., 2011; Anderson et al., 2014; Guse et al., 2014; Dai et al., 2017). Through a sensitivity analysis study on SWAT modeling (Jiang et al., in review), mutual information (MI; Cover and Thomas (2006)) has shown the promise of using a few hundred model realizations to provide similar sensitivity results as the popular Sobol sensitivity analysis (Sobol, 2001) that usually relies on several thousand realizations. As a result, MI is well suited to unravel the relation between model parameters and responses given a few hundred realizations of a state-of-the-art fully-integrated hydrological model.

This study aims to develop a novel knowledge-informed DL method for model calibration by using a few hundred realizations. We leverage MI-based global sensitivity analysis to uncover the dependencies between parameters and observed responses, which are then used to guide the selection of crucial responses as the inputs of DL-assisted inverse mapping. We applied this method in estimating multiple parameters of a fully integrated hydrological model, ATS (Coon et al., 2019), at the Coal Creek watershed, a snow-dominated alpine basin located in Colorado, US. Multiple water years of hydrological observations are used in the ATS model calibration. We further performed a multi-year analysis to investigate the significance of wet and dry years in model calibration. Our study highlights the importance of domain knowledge in uncovering the dependencies among variables of interest before hydrological model calibration.
2 Methods

2.1 Study site

The Coal Creek watershed is located in the western part of the larger East Taylor Watershed in Colorado (Figure 1(a)). The majority of the discharge flows through Coal Creek from the west to the east. The watershed is a HUC12 (Hydrologic Unit Code) watershed encompassing around 53.2 km$^2$ of the drainage area (HUC12 ID, 140200010204). According to the Köppen classification system (Köppen and Geiger, 1930), this high alpine watershed is classified as warm summer and humid continental climate with a significant snow process dominating the hydrological cycle. Based on the long-term Daymet forcing dataset (Thornton et al., 2021), the watershed receives $\sim$530 mm of snowfall annually, dominating its annual precipitation ($\sim$850 mm). This watershed exhibits strong variations in topography with elevations ranging from 2706m to 3770m, where the primary land covers are evergreen forest (62.6%) and shrub (20.5%). Hydrological observations are available through (1) a USGS gaging station (station number 09111250) that records continuous discharge (Q) observations at the watershed outlet; and (2) a remote sensing product of the Moderate Resolution Imaging Spectroradiometer (MODIS) 8-day composite evapotranspiration (ET) at a 500 m resolution. Figure 1 shows the time series of Q and watershed-averaged MODIS ET during the water years from 2016 to 2019, which are used as observations for calibrating the ATS model.

2.2 ATS model setup

ATS is a fully distributed hydrologic model that integrates surface and subsurface flow dynamics (Coon et al., 2019). The surface hydrological process is characterized by a two-dimensional diffusion wave approximation of the Saint-Venant equation. A three-dimensional Richards equation is used to represent the subsurface flow. The model adopts the Priestley Taylor equation to simulate evapotranspiration (ET) from various processes (e.g., snow and plant transpiration), which are coupled with the surface-subsurface hydrological cycle.

We leveraged an existing ATS setup at the Coal Creek watershed (Shuai et al., 2022). The Watershed Workflow package (Coon and Shuai, 2021) was used to delineate the mesh, the surface land covers, and the subsurface characteristics of the watershed. The resulting mesh consists of 171760 cells, formed by a two-dimensional triangle surface mesh followed by 19 terrain-following subsurface layers (Figure 1(a)). The surface mesh contains 8588 triangular cells with varying sizes that range from $\sim$5000m$^2$ near the stream network to $\sim$50000m$^2$ away from the stream network. On the surface, the National Land Cover Database was used to delineate the land cover types. In the subsurface, the 19 layers add up to 28m and contain: (1) 6 soil layers in the top 2m, (2) 12 geological layers in the middle, and (3) 1 bedrock layer in the bottom of the simulation domain. The maximum depth to bedrock (28m) was determined by SoilGrids (Shangguan et al., 2017). The subsurface characteristics of the soil and geological layers are retrieved from the National Resources Conservation Service (NRCS) Soil Survey Geographic (SSURGO) soils database and GLobal HYdrogeology MaPS (GLHYMPS) 2.0 (Huscroft et al., 2018), respectively. The k-means clustering algorithm (Likas et al., 2003) was used to group the soil and geological types based on permeability values, which leads to five soil types and four geological types (Figures 1(c) and (d)). 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 (2) a warm spinup that performed a transient simulation for 10 years under the Daymet forcing. Please refer to Shuai et al. (2022) for the detailed model setup and spinup.

We select a preliminary set of 14 model parameters to be calibrated, which can be categorized into ET, snow, river channel, and subsurface characteristics. The ET parameters include two coefficients used by the Priestley Taylor equation (Priestley and Taylor, 1972) in calculating the potential ET of snow and transpiration, respectively (i.e., priestley_taylor_alpha-snow and priestley_taylor_alpha-transpiration). The snow parameters are the snow melting rate (snowmelt_rate) and the temperature determining the snow melting (snowmelt_degree_diff). The river channel characteristic is the manning’s coefficient (manning_n), which describes the roughness of the surface channel. The subsurface characteristics include the major soil and geological permeability (i.e., perm_s1, perm_s2, perm_s3, perm_s4, perm_s5, perm_g1, perm_g2, perm_g3, and perm_g4). A detailed description of these parameters can be found in Table A1.
2.3 Knowledge-informed model calibration using deep learning

We develop a new methodology to calibrate ATS using knowledge-informed DL, as shown in Figure 2. The key idea is to leverage a data-driven approach to identify the sensitive model response as the inputs to the DL-assisted inverse mapping for each parameter. Here, we use the MI as the sensitivity analysis tool due to its capability to uncover the nonlinear relationships. Derived from Shannon’s entropy (Cover and Thomas, 2006), MI quantifies the shared information between two variables: a model response $Y$ and a model parameter $X$ as follows:

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{X=x} \sum_{Y=y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right),$$

where $p$ is the probability density function and can be estimated by the fixed binning method; and $H(Y) = -\sum_{Y=y} p(y) \log(p(y))$ is Shannon’s entropy. Jiang et al. (in review) show that MI computed by a few hundred realizations with a statistical significance test (SST) can yield comparable sensitivity results with the full Sobol sensitivity analysis that usually uses thousands of realizations through a multivariate sensitivity analysis of SWAT. Therefore, MI is an ideal tool to perform the sensitivity analysis on the several hundred realizations, which are relatively affordable by the computationally-intense ATS model (Cromwell et al., 2021). In this study, we follow a similar strategy of Jiang et al. (in review) 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.
By using MI-based sensitivity analysis, our calibration method involves the following four steps (see Figure 2): (1) narrowing down the parameters to be calibrated using a preliminary MI analysis; (2) computing the parameter sensitivity using a full MI analysis; (3) developing knowledge-guided DL inverse mapping; and (4) parameter estimation with observational error. The details of each step are as follows:

**Step 1: Narrowing down the parameters to be calibrated.** The objective of this step is to further reduce the computational requirement by identifying the parameters that are most relevant to the responses through a coarse-resolution sensitivity analysis. To this end, we first perform a preliminary MI analysis on a sizeable preliminary parameter set. MI is computed between the model response at each time step and each parameter based on Eq.(1). Rather than getting an accurate sensitivity result, this preliminary analysis aims to provide an overview of parameter sensitivity and thus is performed on a small number of realizations to save computational cost. At Coal Creek, we generated 50 realizations based on a total of 14 parameters listed in Table A1 and performed the corresponding ATS runs to compute MI. Sobol sequence (Sobol, 1967), a quasi-Monte Carlo method, is used to generate the parameter ensemble to ensure uniform coverage of the parameter ranges. This preliminary MI analysis would allow filtering out the parameters that show little sensitivities to the model responses, thus reducing the calibrated parameters’ size.

**Step 2: Computing the parameter sensitivity.** We then conduct a complete MI-based sensitivity analysis on the prescreened parameters in Step 1 by using a more significant number of realizations. The analysis accurately quantifies the sensitivity of model responses to each parameter, thereby uncovering the dependency between the two. The analysis also verifies whether the parameters prescreened in Step 1 are ‘truly’ sensitive and can be further excluded if insensitive. At this study site, we generated 400 ATS realizations by sampling from a reduced parameter space at Step 1 using the Sobol sequence, among which 396 ATS runs complete and are used for computing MI.

**Step 3: Developing knowledge-informed DL inverse mapping.** In this step, we develop the DL-based inverse mapping by leveraging the knowledge about the sensitivities between model responses and parameters, learned from the complete MI analysis in Step 2. While the original inverse mapping directly estimates the model parameters from the model responses using a tuned DL model (Figure 3(a)), the complete sensitivity analysis in Step 2 enables the selection of essential responses that are sensitive to a given parameter with non-zero MI as the inputs of the inverse mapping shown in Figure 3(b). Consequently, this ‘filtering’ step is expected to improve the performance of inverse mapping in parameter estimation by reducing the impact of noisy or unimportant model responses.

We developed knowledge-informed inverse mapping for each parameter separately (Figure 3(b)). To evaluate its performance, we also trained the original inverse mapping that estimated all parameters from all the responses by following the same strategy in Cromwell et al. (2021) and Mudunuru et al. (2021) (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. Here, we adopt the
Step 4: Parameter estimation with observation errors. Once an inverse mapping is trained, we first estimate the parameter(s) using the corresponding observations and then use the estimated parameters to perform an ATS simulation. We evaluate the performances of different inverse mappings by comparing the simulated responses with the observations. We further perform uncertainty analysis to analyze how sensitive the DL-based parameter estimation is to the observation errors. To this end, we generate 100 realizations of noisy observations, denoted as $o_n$, such that $o_n = o + \epsilon \times o \times r$, where $o$ is the vector of the original observations, $r$ is the random vector with the same size as $o$, and $\epsilon$ is the standard deviation of the noise and is usually taken as $1/3$ of the observation error. Given an inverse mapping, these noisy observations are used to estimate the ensemble parameter estimations and perform the corresponding ATS ensemble simulation.

3 Results and Discussions

In this section, we present the ATS calibration result at the Coal Creek watershed using the proposed knowledge-informed DL described in Sec. 2.3. We first demonstrate the MI-based sensitivity analysis result, which is facilitated by a preliminary

Figure 3. Illustration of two inverse mappings. While inverse mapping (a) estimates all parameters from all observed responses using a deep learning model (e.g., fully-connected neural net) (Cromwell et al., 2021; Mudunuru et al., 2021), the proposed knowledge-informed inverse mapping (b) estimates each parameter using only responses that share significant MI with the parameter.

multilayer perception (MLP) model as the neural network of inverse mapping, with the details of model configuration tuning and training described in Appendix A.
Table 1: All inverse mappings with and without being knowledge informed developed for calibrating the Advanced Terrestrial Simulator (ATS) at the Coal Creek watershed.

<table>
<thead>
<tr>
<th>Inverse mapping</th>
<th>Response input</th>
<th>Input years</th>
<th>Guided by MI result?</th>
</tr>
</thead>
<tbody>
<tr>
<td>qonly-3yrs</td>
<td>Q</td>
<td>WY2016-2018</td>
<td>No</td>
</tr>
<tr>
<td>mi-qonly-3yrs</td>
<td>Q</td>
<td>WY2016-2018</td>
<td>Yes</td>
</tr>
<tr>
<td>etonly-3yrs</td>
<td>ET</td>
<td>WY2016-2018</td>
<td>No</td>
</tr>
<tr>
<td>mi-etonly-3yrs</td>
<td>ET</td>
<td>WY2016-2018</td>
<td>Yes</td>
</tr>
<tr>
<td>qet-3yrs</td>
<td>Q, ET</td>
<td>WY2016-2018</td>
<td>No</td>
</tr>
<tr>
<td>mi-qet-3yrs</td>
<td>Q, ET</td>
<td>WY2016-2018</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-1yr-1</td>
<td>Q</td>
<td>WY2016</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-1yr-2</td>
<td>Q</td>
<td>WY2017</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-1yr-3</td>
<td>Q</td>
<td>WY2018</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-2yrs-12</td>
<td>Q</td>
<td>WY2016/2017</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-2yrs-13</td>
<td>Q</td>
<td>WY2016/2018</td>
<td>Yes</td>
</tr>
<tr>
<td>mi-qonly-2yrs-23</td>
<td>Q</td>
<td>WY2017/2018</td>
<td>Yes</td>
</tr>
</tbody>
</table>

analysis using fewer realizations to narrow down the parameters to be calibrated. Then, we report the result of the trained DL-assisted inverse mapping performance using all three years of simulations on the test dataset. By using Q alone, our result shows the improved ATS simulation calibrated by knowledge-informed DL inverse mapping over the traditional mapping. However, we also identify the ET observation extrapolation issue that impedes the calibration performance. Last, we present the multi-year analysis using observed Q only to estimate the parameters regarding (1) the performance of the simulated Q driven by the estimated parameters; and (2) the impact of the observational error on the parameter estimations using inverse mapping.

3.1 MI-based sensitivity analysis

Parameter prescreening using a preliminary MI analysis. We calculated the MI for each daily Q/ET with each of the parameters listed in Table A1. Figure 4 shows the estimation of MI by using the 50 ATS runs in the water year 2016 (note that 20 ATS runs didn’t complete after the end of the water year 2016). We then ranked the parameters based on the temporally-averaged MI across the analyzed time period, as plotted on the right panel of each MI heatmap. The MI heatmaps show the varying sensitivities of different parameters over time. For instance, while Q (Figure 4(a)) is mainly sensitive to the permeability (e.g., perm_g1) during low flow seasons, snow melting parameters (e.g., snowmelt_degree_diff) play more critical roles in high flow periods than others. On the other hand, ET (Figure 4(b)) is mostly controlled by the two Priestley Taylor alpha coefficients in low and high flows, respectively. Based on the temporally-averaged MI, we find that Q is mostly sensitive to perm_s3, perm_s4, perm_g1, perm_g4, snowmelt_rate, snowmelt_degree_diff, and priestley_taylor_alpha_transpiration, and
Figure 4. A preliminary Mutual Information (MI)-based sensitivity analysis using 50 realizations of the Advanced Terrestrial Simulator (ATS) on (a) the discharge (Q) and (b) evapotranspiration (ET). The top panel of each subplot shows the ATS ensemble (gray) and the corresponding observations (black). (Note that only water year 2016 is completed by all 50 runs).

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).

The full MI analysis. We performed the ATS ensemble simulations on 400 realizations of the reduced parameter set and computed the MI heatmaps on the completed 396 realizations during water years 2016-2019 shown in Figure 5. By using more
Figure 5. A full MI-based sensitivity analysis using 396 realizations of the Advanced Terrestrial Simulator (ATS) on (a) the discharge (Q) and (b) evapotranspiration (ET). The top panel of each subplot shows the ATS ensemble (gray) and the corresponding observations (black).

realizations, this complete MI analysis shows a better delineation of parameter sensitivity than the preliminary analysis due to its convergence on estimation (see the convergence of the parameter rankings in Figure A1). 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. In fact, by comparing the parameter sensitivity rankings of both Q and ET based on the full and preliminary analysis shown in Figure 6, we find that this full analysis (1) improves the MI estimations (e.g., perm_s3); and
Figure 6. Sensitivity rankings based on mutual information (MI) on (a) the discharge (Q) and (b) evapotranspiration (ET) of both the preliminary and full MI analyses.

(2) identifies the insensitive parameters (e.g., perm_s4) that are falsely considered sensitive due to the limited samples in the preliminary analysis. Since both Q and ET share little MI with perm_s4, we exclude it from the parameters to be calibrated. In the following, we leveraged this MI result to identify the sensitive responses (i.e., with non-zero MI) to be used as the inputs of the knowledge-informed DL inverse mapping.

3.2 ATS calibration using 3yr Q/ET/Q-ET

We report the performances of the trained DL inverse mappings, with and without being knowledge-informed using MI-based sensitivity analysis, on the test dataset with regards to the modified Kling-Gupta Efficiency (mKGE, Kling et al., 2012) and the Nash-Sutcliffe Efficiency (NSE), as shown in Figure 7. We use blue and green to represent the performance of the mappings with and without being knowledge informed, respectively. Blank, cross, and circle textures stand for the mapping taking as inputs only discharge (qonly), only evapotranspiration (etonly), and both (qet), respectively. Overall, most inverse mappings are trained well with both mKGE and NSE greater than 0.8, with a few exceptions (e.g., the two snow melting coefficients, perm_g1, and perm_g4 using etonly-3yr and mi-etonly-3yr). The inferior estimations of the four parameters are due to their minimal MI shared with ET (Figure 5) such that ET dynamics are insufficient to inform the two parameters.

It can be observed from Figure 7 that a knowledge-informed inverse mapping (i.e., the blueish/black bars) generally outperforms and has higher mKGE/NSE than its counterpart that uses all the observed time steps as inputs (i.e., the redish bars). Noticeably, the NSE of knowledge-informed mapping increases when estimating the two Priestley-Taylor coefficients, the two snow melting parameters (except that estimated by mi-etonly-3yrs), and perm_s3 regardless of which model response is used in calibration. Besides, the extent of how sensitive the two responses are to a parameter also plays a role in the performance of the inverse mappings. The two Priestley-Taylor coefficients, which are the two most sensitive parameters to ET, are better estimated by using ET than using only Q dynamics. On the other hand, using Q yields superior performance over using ET in
Figure 7. Performance of the developed deep learning (DL) inverse mapping on the test dataset using three years of model responses 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. As a result, when both Q and ET are used, the knowledge-guided inverse mappings (the black bars) turn out to be the best calibration tool for most of the parameters.

ATS forward runs on estimated parameters. We estimated the parameters using each of the inverse mappings (see Table A2) and performed the corresponding ATS forward runs. Figure 8 shows the Q and ET simulations by using the estimated parameters (the blue lines) and the observations (the black lines), with their 1-to-1 scatter plots shown on the side. The parameter estimated by using knowledge-informed inverse mapping improves the calibrated Q simulation. When using only Q for calibration, mKGE increases from 0.65 (qonly-3yrs) to 0.80 (mi-qonly-3yrs).
Figure 8. Forward runs of the Advanced Terrestrial Simulator (ATS) based on the estimated parameters from inverse mappings using all three years’ responses in Table 1. (a-d) are the simulated discharge (Q) using qonly-3yrs, mi-qonly-3yrs, qet-3yrs, and mi-qet-3yrs, respectively. (e-h) are the simulated evapotranspiration (ET) using etonly-3yrs, mi-etonly-3yrs, qet-3yrs, and mi-qet-3yrs, respectively.

The extrapolation issue of ET observations. Meanwhile, ET simulations deteriorate when using knowledge-informed inverse mapping, with NSE decreasing from 0.61 (etonly-3yrs) to 0.53 (mi-etonly-3yrs). This surprising result is attributed to the extrapolation issue of ET observations. 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 (Khan et al., 2018; Xu et al., 2019). Xu et al. (2019) 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 (Senay et al., 2013), particularly during the dry seasons (i.e., May through September) (Xu et al., 2019). 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.

One consequence of these ET ‘outliers’ is the inability of the inverse mapping to reasonably estimate the parameters. The resulting estimations of priestley_taylor_alpha-snow, whose sensitivities with ET mainly occur during the low ET period, greatly surpass its maximum sample threshold (i.e., 1.2; see Table A1) and range from 1.8 to 2.1 (see Table A2). These unreliable parameter estimations make the knowledge guidance less valid. Another evidence of the adverse impact of the extrapolation issue is the inferior simulation on Q when both Q and ET are used in calibration (e.g., qet-3yrs and mi-qet-3yrs). While using knowledge-informed inverse mapping still increases the NSE to 0.47 (mi-qet-3yrs) from 0.19 (qet-3yrs), both are worse than that of calibrating against Q alone (i.e., qonly-3yrs and mi-qonly-3yrs). This extrapolation issue underscores the significance of defining the acceptable parameter sampling ranges and understanding the observation uncertainty before calibration. We thus performed the remaining analysis using only Q for calibration to avoid the ET extrapolation impact.

3.3 ATS calibration using multi-years Q

To investigate how wet/dry water years impact the model calibration, we further employed the knowledge-informed inverse mapping to calibrate ATS on different numbers of years of Q ensemble. To this end, we used one year, two years, and all three years of Q to develop the inverse mapping for model calibration. Figure 9 plots the calibrated ATS forward runs of Q, along with the observations. First, we find that increasing the number of observed years in calibration does not necessarily improve the performance of ATS simulation on discharge. This is, in fact, consistent with the performance of the corresponding inverse mapping on the test dataset (see Figure A2 in the appendix). In other words, using only one year of observations can yield a similar calibration result with using multiple years at this watershed. This can be attributed to the similar seasonal cycle of the whole-year discharge dynamics such that multi-year dynamics do not necessarily add more information to improve the calibration. This multi-year analysis underscores the potential of using fewer years than the number of available observed years in model calibration, which can save the computational time for one ensemble run and lead to more ensemble simulations given a fixed computational budget.

The significance of dry year dynamics. Despite the similar calibration results between using one year and multiple years, we find that by including the dry year (i.e., the 2nd year), the calibration greatly improved the simulation of Q over the scenarios using only wet years dynamics. Using only the dry year (i.e., 2nd year or mi-qonly-1yr-2) in calibration generates the best-simulated Q with NSE=0.83 and mKGE=0.82. It outperforms the simulated Q using either of the other two wet years (i.e., mi-qonly-1yr-1 and mi-qonly-1yr-3) with NSE=0.74 and 0.67 (mKGE=0.79 and 0.73), respectively. Particularly, mi-qonly-1yr-3, that uses the 3rd and wettest year in calibration, generates the worst simulation of Q among the three inverse mappings using one-year dynamics. For the inverse mappings using two years, the simulated Q of the mappings including the dry year (i.e., mi-qonly-2yrs-12 and mi-qonly-2yrs-23) are better than that of using the two wet years (i.e., mi-qonly-2yrs-13). We also observe that mi-qonly-2yrs-12 (NSE=0.85 and mKGE=0.78) outperforms mi-qonly-2yrs-23 (NSE=0.72 and mKGE=0.66) in the simulation of Q, which probably results from the better performance of using the 1st year (mi-qonly-1yr-1) only than
Figure 9. Forward runs of the Advanced Terrestrial Simulator (ATS) based on the estimated parameters from knowledge-informed inverse mappings using multiple years of discharge (Q) in Table 1. (a-c) are the simulated discharge (Q) using each of the three water years to calibrate (i.e., mi-qonly-1yr-1, mi-qonly-1yr-2, and mi-qonly-1yr-3). (d-f) are the simulated discharge (Q) using two of the three water years to calibrate (i.e., mi-qonly-2yrs-12, mi-qonly-2yrs-23, and mi-qonly-2yrs-13). (g) is the simulated discharge using all three years to calibrate (i.e., mi-qonly-3yrs).

using the wetter 3rd year (mi-qonly-1yr-3). And using all three years (i.e., mi-qonly-3yrs) also guarantees reasonably well simulations of Q with NSE=0.80 and mKGE=0.80.

The enhanced calibration performance by including the dry year dynamics is probably due to the improved dependencies between discharge and subsurface features during low flow periods. This can be observed in the yearly averaged MI bar plots in Figure 10. With reduced sensitivities on the snow parameters, the dry year shows that the dominant permeability in the soil layer (i.e., perm_s3) proves to be the most sensitive parameter to discharge (with averaged MI 0.2). Meanwhile, the averaged MI of perm_s3 decreases to only around 0.1 in the 3rd year by using which the calibrated Q is the worst with NSE=0.67 and mKGE=0.73. In fact, all the inverse mappings using the dry year consistently estimates higher perm_s3 (i.e., 10.9 log_{10}(m^2)) than the other estimates ranging from -11.4 to -12.2 log_{10}(m^2) (see Table A3 in the appendix). The higher soil permeability...
estimated by the dry year has a better capability in draining the surface water and thus reduces the outlet discharge spikes during late spring and early autumn, thus yielding the simulations of $Q$ more consistent with the observations plotted in Figure 9.

Our finding on the significance of dry year discharge in model calibration indirectly supports some recent studies. Pool et al. (2019) found that high flow provides limited information to calibrate models in snow-dominated catchments. This is mainly because there are less discharge fluctuations during snow melting or high flow period than rainfall-fed catchments (Viviroli and Seibert, 2015). The decreased role of high flow, in turn, enhance the importance of low flow period in calibration, particularly in dry years. Indeed, in this watershed, while there are limited diurnal fluctuations during snow melting period through all the three years, we do observe stronger diurnal discharge fluctuations during summer and autumn of the dry year (i.e., WY2017) than the other two wetter years (see Figure A3 in the appendix).

**Impact of observation errors.** For each of the seven inverse mappings, we set 5% observational error and generated 100 sets of estimated parameters by using 100 realizations of noisy observed discharges. Figure 11 shows the barplots of these ensemble estimations of parameters. Overall, most estimated parameters (i.e., priestley_taylor_alpha-transpiration, snowmelt_degree_diff, perm_s3, and perm_g1) show little variability, indicating the robustness of the trained inverse mappings. Although there is higher variability for the other parameters (i.e., priestley_taylor_alpha-snow, snowmelt_rate, and perm_g4), these parameters share less MI with $Q$ than the others (Figure 10). Therefore, the larger variations of these less sensitive parameters have little impact on the corresponding ATS forward runs. Indeed, as shown in Figure 12, the ATS ensemble runs using inverse mapping mi-qonly-3yrs demonstrate little variations of the ensemble runs of the discharges compared with the corresponding run without observational error, which verifies the robustness of these inverse mappings against the observational error.
Figure 11. Box plots of ensemble parameters estimated knowledge-informed inverse mappings of multiple years of discharge (Q) using 100 realizations of noisy discharge observations with 5% observation error, categorized into (a) evapotranspiration (ET) parameters; (b) snow parameters; and (c) permeability parameters.

4 Conclusions

We developed a novel model calibration methodology that leverages MI-based sensitivity analysis to guide the architecture of the DL-assisted inverse mapping based on only a few hundred realizations. A case study in the Coal Creek watershed shows that the calibrated ATS using such knowledge-informed DL simulates the discharge well with both mKGE and NSE up to 0.8, outperforming its counterpart that does not adopt the knowledge guidance and only achieves NSE and mKGE around 0.54 and 0.65, respectively (Figures 8(a) and (b)).
Nevertheless, we observe the adverse impact of the observation outliers when calibrating ATS against remote sensing ET product (Figures 8(e)-(h)). These outliers deteriorate the performance of the inverse mapping in parameter estimation, thereby worsening the calibrated ATS forward runs. Based on the inverse mappings taking ET as inputs, the estimated priestley_taylor_alpha-snow, which is the most sensitive to ET (Figure 5(b)), greatly exceeds the range of the parameter sample range (Tables A1 and A2). While the uncertainty of MODIS ET (Khan et al., 2018; Xu et al., 2019) also contributes to this extrapolation issue, this result underscores the significance of suitably defining the parameter sample range to assure the uncertainty of ensemble simulations covers the observations.

We further find that using one or two years of observations in calibration yields similar or even better results than that of using three years. This encouraging result highlights the importance of including abnormal year data for model calibration. This would significantly reduce the computational cost of each model run, increasing the number of model ensemble realizations used in developing calibration techniques. While earlier studies found that several years of discharge observations are necessary to achieve a reasonable model calibration (Sorooshian et al., 1983; Yapo et al., 1996; Perrin et al., 2007) by using either a semi-distributed or a bucket model, we suspect that employing a fully integrated and high resolution model like ATS can greatly reduce the errors due to the improved physical representation in the model, thus requiring longer observation periods for calibration. Hence, future work can focus on developing a systematical approach in identifying the most ‘important’ observed period in model calibration.

The recently emerged knowledge-guided DL is swiftly gaining popularity in earth science. This study demonstrates one of its applications in calibrating a computationally expensive hydrological model. The developed methodology can be readily adopted in other watersheds to calibrate other models. With the increasing complexity of earth system models, we believe such knowledge-guided DL calibration can pave the way for efficient yet effective model calibration without increasing significant computational demand.
Appendix A: Development of deep learning-assisted inverse mapping and hyperparameter tuning

Each of the inverse mappings listed in Table 1 is developed using a multilayer perception (MLP) model as follows. Let’s denote the number of inputs, outputs, and hidden layers as $N_i$, $N_o$, and $N_l$, respectively. $N_i$ depends on the selections of the response variables and the number of years, which ranges from $\sim 100$ 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 seven (i.e., estimating all parameters by 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 $N_{h,l} = \lfloor \frac{N_i - N_o}{N_l} \times l \rfloor$ (where $1 \leq l \leq N_l$ and $\lfloor \cdot \rfloor$ is the floor function). In doing so, the information from the long time series of observed responses can be gradually propagated to the estimated parameters with limited numbers (i.e., 1 or 7), and at the same time, only $N_l$ needs to be tuned instead of both $N_l$ and $N_{h,l}$. We split the 396 realizations into 300/50/46 for training/validation/test, respectively. Using the Adam optimization algorithm, 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 algorithm (Kingma and Ba, 2014) was used to optimize the loss. We tuned each MLP by varying $N_l = [1, 3, 5, 7, 9, 10]$ and the learning rate $l_r = [1e^{-5}, 1e^{-4}, 1e^{-3}]$. 

https://doi.org/10.5194/hess-2022-282
Preprint. Discussion started: 5 August 2022
© Author(s) 2022. CC BY 4.0 License.
Figure A1. Convergence of the parameter sensitivity rankings by using the averaged mutual information (MI) on discharge (Q) and evapotranspiration (ET) using 30, 50, 100, 200, 300, and 396 model realizations.
Figure A2. Performance of the developed deep learning (DL) inverse mapping on the test dataset using multi-years discharges (Q) on the test dataset with regards to (a) the modified Kling-Gupta Efficiency (mKGE) and (b) the Nash-Sutcliffe Efficiency (NSE).
Figure A3. Logrithmatic discharge observation from water year 2016 through water year 2018 at the USGS gage station 09111250. The top panel shows the corresponding snow (yellow) and rain (blue) precipitation.
Table A1: List of the parameters used in the preliminary or full mutual information analysis.

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Description [unit]</th>
<th>Min</th>
<th>Max</th>
<th>Used in Full MI analysis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>perm_s1</td>
<td>permeability of the soil layer s1 [log m$^2$]</td>
<td>-12.548</td>
<td>-10.548</td>
<td>No</td>
</tr>
<tr>
<td>perm_s2</td>
<td>permeability of the soil layer s2 [log m$^2$]</td>
<td>-13.022</td>
<td>-11.022</td>
<td>No</td>
</tr>
<tr>
<td>perm_s3</td>
<td>permeability of the soil layer s3 [log m$^2$]</td>
<td>-12.710</td>
<td>-10.710</td>
<td>Yes</td>
</tr>
<tr>
<td>perm_s4</td>
<td>permeability of the soil layer s4 [log m$^2$]</td>
<td>-12.324</td>
<td>-10.324</td>
<td>Yes</td>
</tr>
<tr>
<td>perm_s5</td>
<td>permeability of the soil layer s5 [log m$^2$]</td>
<td>-13.688</td>
<td>-11.688</td>
<td>No</td>
</tr>
<tr>
<td>perm_g1</td>
<td>permeability of the geological layer g1 [log m$^2$]</td>
<td>-13.500</td>
<td>-11.500</td>
<td>Yes</td>
</tr>
<tr>
<td>perm_g2</td>
<td>permeability of the geological layer g2 [log m$^2$]</td>
<td>-11.520</td>
<td>-9.520</td>
<td>No</td>
</tr>
<tr>
<td>perm_g3</td>
<td>permeability of the geological layer g3 [log m$^2$]</td>
<td>-16.200</td>
<td>-14.200</td>
<td>No</td>
</tr>
<tr>
<td>perm_g4</td>
<td>permeability of the geological layer g4 [log m$^2$]</td>
<td>-14.000</td>
<td>-12.000</td>
<td>Yes</td>
</tr>
<tr>
<td>snowmelt_rate</td>
<td>snow melt rate [mm day$^{-1}$ C$^{-1}$]</td>
<td>1.37</td>
<td>5.48</td>
<td>Yes</td>
</tr>
<tr>
<td>snowmelt_degree_diff</td>
<td>air-snow temperature difference [deg C]</td>
<td>1</td>
<td>5</td>
<td>Yes</td>
</tr>
<tr>
<td>manning_n</td>
<td>Manning’s coefficient [-]</td>
<td></td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>priestley_taylor_alpha-snow</td>
<td>Priestley Taylor coefficient of canopy transpiration [-]</td>
<td>0.3</td>
<td>1.2</td>
<td>Yes</td>
</tr>
<tr>
<td>priestley_taylor_alpha-transpiration</td>
<td>Priestley Taylor coefficient of snow evaporation [-]</td>
<td>0.3</td>
<td>1.2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table A2: Estimated parameters from the inverse mappings using three years of Q/ET/Q and ET listed in Table 1 (Note that perm_g1 is not estimated by mi-etonly-3yrs because ET is not sensitive to perm_g1 at the analyzed three water years shown in Figure 5).

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>qonly-3yrs</th>
<th>etonly-3yrs</th>
<th>qet-3yrs</th>
<th>mi-qonly-3yrs</th>
<th>mi-etonly-3yrs</th>
<th>mi-qet-3yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>perm_s3</td>
<td>-11.859</td>
<td>-12.720</td>
<td>-12.715</td>
<td>-10.992</td>
<td>-12.709</td>
<td>-12.708</td>
</tr>
<tr>
<td>perm_g1</td>
<td>-11.790</td>
<td>-10.906</td>
<td>-10.722</td>
<td>-10.451</td>
<td>n/a</td>
<td>-10.451</td>
</tr>
<tr>
<td>perm_g4</td>
<td>-12.603</td>
<td>-10.150</td>
<td>-13.514</td>
<td>-12.536</td>
<td>-12.791</td>
<td>-12.276</td>
</tr>
<tr>
<td>snowmelt_rate</td>
<td>2.538</td>
<td>4.289</td>
<td>2.880</td>
<td>2.569</td>
<td>2.660</td>
<td>2.852</td>
</tr>
<tr>
<td>snowmelt_degree_diff</td>
<td>1.697</td>
<td>7.755</td>
<td>1.822</td>
<td>2.324</td>
<td>2.422</td>
<td>2.365</td>
</tr>
<tr>
<td>priestley_taylor_alpha-snow</td>
<td>0.296</td>
<td>2.130</td>
<td>2.135</td>
<td>0.354</td>
<td>1.925</td>
<td>1.834</td>
</tr>
<tr>
<td>priestley_taylor_alpha-transpiration</td>
<td>2.018</td>
<td>0.665</td>
<td>0.746</td>
<td>1.564</td>
<td>0.515</td>
<td>0.846</td>
</tr>
</tbody>
</table>
Table A3: Estimated parameters from the inverse mappings using multi-years of Q listed in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Knowledge-informed inverse mappings only using discharge (i.e., mi-qonly-*)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1yr-1</td>
</tr>
<tr>
<td>perm_g1</td>
<td>-10.917</td>
</tr>
<tr>
<td>snowmelt_rate</td>
<td>2.308</td>
</tr>
<tr>
<td>snowmelt_degree_diff</td>
<td>2.149</td>
</tr>
<tr>
<td>priestley_taylor_alpha-snow</td>
<td>0.450</td>
</tr>
<tr>
<td>priestley_taylor_alpha-transpiration</td>
<td>1.541</td>
</tr>
</tbody>
</table>

**Author contributions.** PJ, PS, and XC designed the numerical experiment. PJ developed and trained the deep learning-assisted inverse mapping. PS performed the ATS simulations. PJ prepared the manuscripts with contributions from all co-authors. All authors provided critical feedback and inputs to the manuscript.

**Competing interests.** The contact author has declared that neither they nor their co-authors have any competing interests.

**Disclaimer.** Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Code and data availability.** The datasets and scripts used in this study will be made public upon the publication of this manuscript.

**Acknowledgements.** This work was funded by the ExaSheds project, which was supported by the United States Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Systems Sciences Division, Data Management Program, under Award Number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the United States Department of Energy under contract DE-AC02-05CH11231. Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the United States Department of Energy or the United States Government. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish, or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results.
of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan, last access: 20 May 2022).
References


