The Journal of Open Source Software

KIM: Knowledge-Informed Mapping (KIM) Toolkit

Peishi Jiang[1,2], **Aaron Wang**[1], **Susannah M. Burrows**[1], **Naser Mahfouz**[1], **and Xingyuan Chen**[1]

1 Atmospheric, Climate, and Earth Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA **2** Civil, Construction and Environmental Engineering, University of Alabama, Tuscaloosa, Alabama, USA

Peishi Jiang[1,2], **Aaron Wang**[1], **Susannah M. Burrows**[1], **Naser Mahfouz**[1], **Xingyuan Chen**[1]

[1] Atmospheric, Climate, and Earth Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA
[2] Civil, Construction and Environmental Engineering, University of Alabama, Tuscaloosa, AL, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Vangelis Kourlitis

Submitted: 21 August 2025
Published: unpublished

Summary

We present a Knowledge-Informed Mapping toolkit in Python programming language, named KIM, to optimize the development of the mapping from a vector of inputs $\mathbf{X}$ to a vector of outputs $\mathbf{Y}$. KIM builds on the methodology development of deep learning-based inverse mapping in Jiang et al. (2023) and A. Wang et al. (2025). KIM offers a preliminary understanding of data interdependencies while optimizing the training step with uncertainty accounted for. We expect this toolkit will be helpful to glue the model data integration for Earth science applications.

Statement of need

Striving for scientific hypothesis testing and discovery, Earth scientists oftentimes develop data-driven mappings – either for inverse modeling, as part of model calibration, or forward modeling, as an emulator. Both approaches benefit from an efficient way of mapping, , that projects from a vector of inputs $\mathbf{X}$ to a vector of outputs $\mathbf{Y}$. Such mapping approach has seen successes in addressing inverse and forward problems in multiple studies across Earth sciences (Cromwell et al., 2021; HU et al., 2014; Krasnopolsky & Schiller, 2003; Mudunuru et al., 2022).

Nevertheless, constructing the mapping that connects all inputs $\mathbf{X}$ to all outputs $\mathbf{Y}$ is usually challenging due to (1) limited data/simulations for training; (2) uninformative relations between some members of $\mathbf{X}$ and $\mathbf{Y}$; and (3) the structural uncertainty of the mapping . To that, Jiang et al. (2023) and A. Wang et al. (2025) leveraged the idea of integrating scientific knowledge with deep learning (Willard et al., 2022) to develop knowledge-informed mapping (KIM). The goal of this paper is to document and open source KIM for a general public usage. Figure 1 shows the general procedures of KIM which are detailed in the next section.

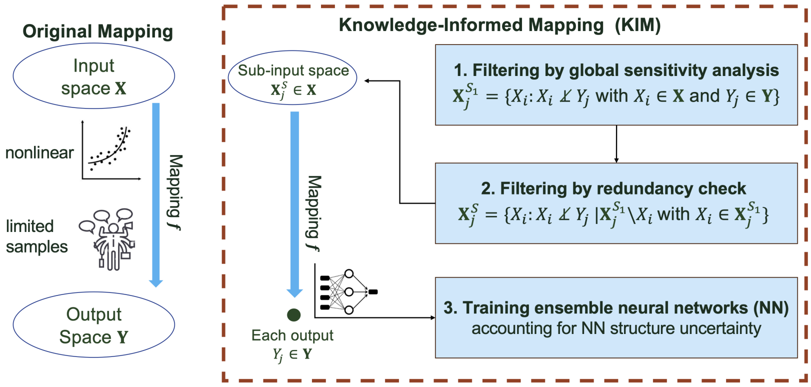

Figure 1: Comparison between KIM and the original mapping.

Mathematical approach

Consider a vector of inputs $\mathbf{X} = [X_1, ..., X_{N_x}]$ and a vector of outputs $\mathbf{Y} = [Y_1, ..., Y_{N_y}]$. The objective is to build up a mapping function f from $\mathbf{X}$ to $\mathbf{Y}$, such that $f : \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$, based on N_e pairs/realizations of $\mathbf{X}$ and $\mathbf{Y}$. Instead of developing a lumped mapping, we aim to develop a separate inverse mapping f_i for each $Y_j \in \mathbf{Y}$ by using a reduced space $\mathbf{X}_j^S \in \mathbf{X}$ that is most relevant to Y_j, such that $f_j : \mathbb{R}^{N_{x_j}} \to \mathbb{R}$ (see examples in Jiang et al. (2023) and A. Wang et al. (2025)), which involves the following steps.

Step 1: Filtering by global sensitivity analysis. We first perform a mutual information-based global sensitivity analysis to narrow down a subset $\mathbf{X}_j^{S_1}$, each of which shares zero information with Y_j such that:

$$\mathbf{X}_j^{S_1} = \{X_i : I(X_i; Y_j) \neq 0 \quad \text{with } X_i \in \mathbf{X}\},$$

where $I(X_i; Y_j)$ is the mutual information between X_i and Y_j (Cover & Thomas, 2006). Based on the N_e realizations, I is calculated on the joint probability of X_i and Y_j using either binning method or k-nearest-neighbor method.

Step 2: Filtering by redundancy check. Then, we conduct a further assessment that filters out any model output in $\mathbf{X}_j^{S_1}$ whose dynamics are redundant to Y_j given the knowledge of other outputs. This is achieved through a conditional independence test using conditional mutual information (Cover & Thomas, 2006) given as:

$$\mathbf{X}_j^{S} = \{X_i : I(X_i; Y_j | \mathbf{X}_j^{S_1} \backslash X_i) \neq 0 \quad \text{with } X_i \in \mathbf{X}_j^{S_1}\},$$

where $\mathbf{X}_j^{S_1} \backslash X_i$ is the remaining set of $\mathbf{X}_j^{S_1}$ by excluding X_i; $I(X_i; Y_j | \mathbf{X}_j^{S_1} \backslash X_i)$ is the conditional mutual information between X_i and Y_j conditioning on $\mathbf{X}_j^{S_1} \backslash X_i$. $I(X_i; Y_j | \mathbf{X}_j^{S_1} \backslash X_i) = 0$ indicates that X_i and Y_j are independent given the knowledge of $\mathbf{X}_j^{S_1} \backslash X_i$.

Step 3: Uncertainty aware estimation by training ensemble neural networks. For each parameter Y_i, we train an ensemble of fully-connected neural networks by varying the hyperparameters, including the number of hidden layers, the number of hidden neurons, and the learning rate. We split the N_e model realizations into training, validation, and testing dataset. When evaluating the estimation on the test dataset, we further quantified the bias and uncertainty of the prediction as:

$$\text{Bias} = E(|\mu_w - y|)$$
$$\text{Uncertainty} = E(\sigma_w/|y|),$$

where E is the expectation operator; y is the true value; μ_w and σ_w are the mean and standard deviation of the ensemble predictions weighted by their accuracy in the validation set.

Examples

We present two applications of KIM in performing inverse modeling, with Jupyter notebook provided in the repository to guide the package usage. For each case, we developed three types of inverse mappings: (1) the original inverse mapping without knowledge-informed, denoted as M_0; (2) the knowledge-informed inverse mapping only using global sensitivity analysis (Step 1), denoted as M_1; and (3) the knowledge-informed inverse mapping using both Step 1 and Step 2, denoted as M_2. The configurations can be found in the example jupyter notebook.

Case 1: Calibrating a cloud chamber model. Cloud chamber model has been widely applied as a virtual reality of a true cloud chamber to study turbulence, clouds, and their interactions (Thomas et al., 2019; Aaron Wang, Ovchinnikov, Yang, Schmalfuss, et al., 2024; Aaron Wang, Ovchinnikov, Yang, Cantrell, et al., 2024; Aaron Wang, Krueger, et al., 2024; Aaron Wang et al., 2025). The objective of this example is to estimate two key parameters, i.e., the scaling coefficients of wall fluxes (λ_w) and collision processes (λ_c) using inverse mapping. To that, an ensemble of 513 model runs were generated based on a model set up detailed in A. Wang et al. (2025), by varying the values of the two parameters using Sobol sequence. 27 virtual sensors are configured, each of which 'records' multiple variables including flow properties and cloud properties.

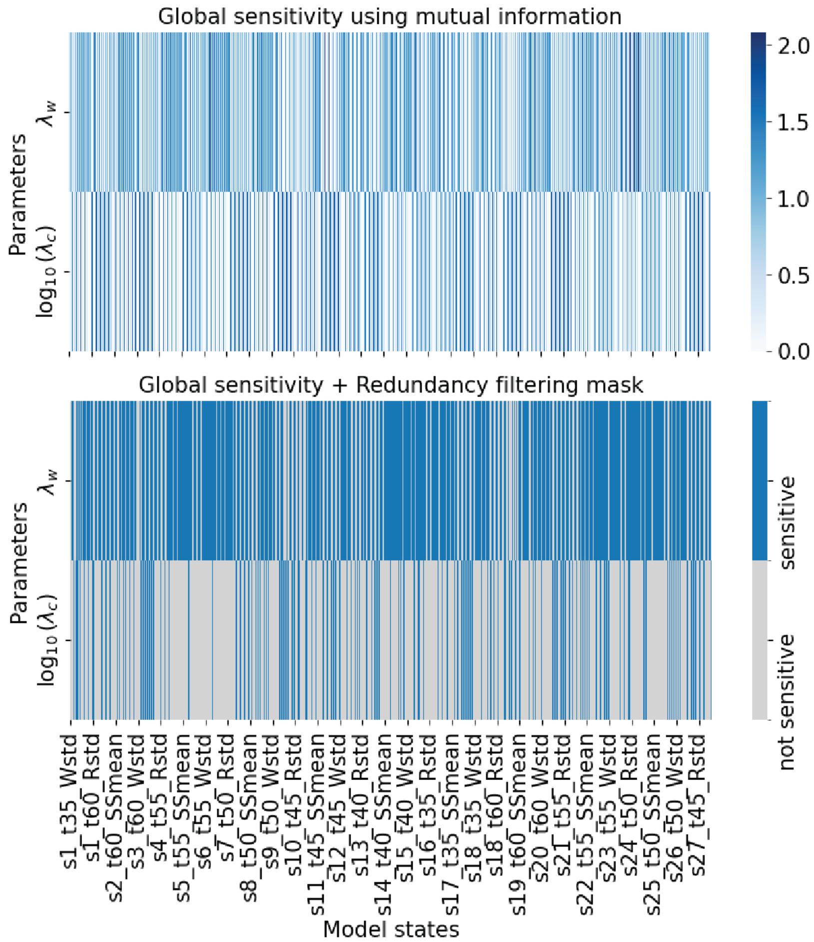

Figure 2: Preliminary analysis of cloud chamber ensemble modeling.

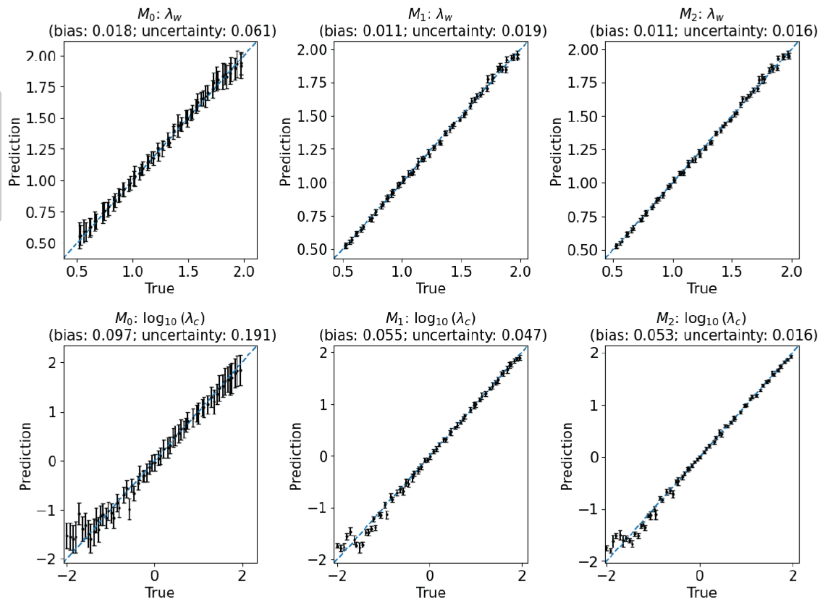

Figure 3: Parameter estimation of the cloud chamber model.

Case 2: Calibrating an integrated hydrological model. The Advanced Terrestrial Simulator (ATS) is an integrated hydrological models used to simulate hydrological fluxes across a watershed (Coon et al., 2019). Here, we calibrated ATS against the streamflow observations at the outlet of Coal Creek watershed, CO, USA. The objective is to estimate eight models parameters categorized into evapotranspiration (ET), snow melting, and subsurface permeability. See Jiang et al. (2023) for more detailed information.

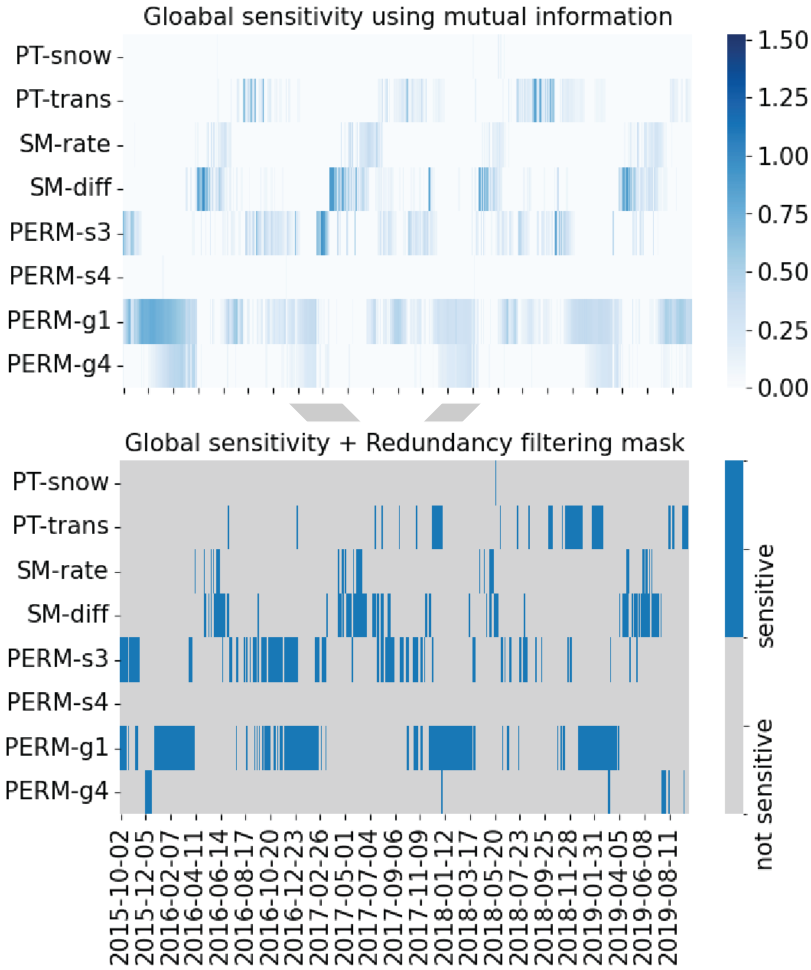

Figure 4: Preliminary analysis of ATS ensemble modeling.

The Journal of Open Source Software

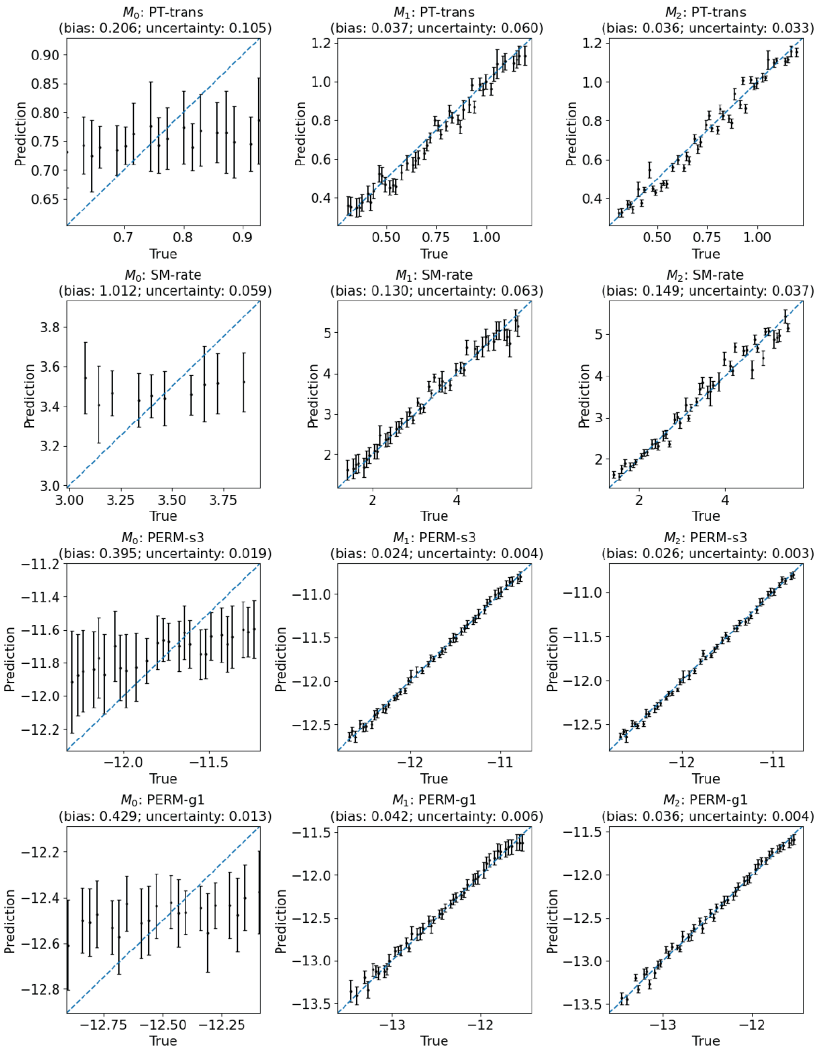

Figure 5: Parameter estimation of the ATS model.

Acknowledgements

This work was supported by both the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory and the IDEAS-Watersheds project. The Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. Pacific Northwest National Laboratory is operated for the DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. The IDEAS-Watersheds project is funded by the U.S. Department of Energy (DOE), Office of Science (SC) Biological and Environmental Research (BER) program, as part of BER's Environmental System Science (ESS) program.

References

Coon, E., Svyatsky, D., Jan, A., Kikinzon, E., Berndt, M., Atchley, A., Harp, D., Manzini, G., Shelef, E., Lipnikov, K., Garimella, R., Xu, C., Moulton, D., Karra, S., Painter, S., Jafarov, E., & Molins, S. (2019). *Advanced terrestrial simulator*. [Computer Software] https://doi.org/10.11578/dc.20190911.1. https://doi.org/10.11578/dc.20190911.1

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory (wiley series in telecommunications and signal processing)*. Wiley-Interscience. ISBN: 0471241954

Cromwell, E., Shuai, P., Jiang, P., Coon, E. T., Painter, S. L., Moulton, J. D., Lin, Y., & Chen, X. (2021). Estimating watershed subsurface permeability from stream discharge data using deep neural networks. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.613011

HU, Y., YU, X., LI, S., CHEN, G., ZHOU, Y., & GAO, Z. (2014). Improving the accuracy of geological model by using seismic forward and inverse techniques. *Petroleum Exploration and Development*, *41*(2), 208–216. https://doi.org/https://doi.org/10.1016/S1876-3804(14)60024-0

Jiang, P., Shuai, P., Sun, A., Mudunuru, M. K., & Chen, X. (2023). Knowledge-informed deep learning for hydrological model calibration: An application to coal creek watershed in colorado. *Hydrology and Earth System Sciences*, *27*(14), 2621–2644. https://doi.org/10.5194/hess-27-2621-2023

Krasnopolsky, V. M., & Schiller, H. (2003). Some neural network applications in environmental sciences. Part i: Forward and inverse problems in geophysical remote measurements. *Neural Networks*, *16*(3), 321–334. https://doi.org/https://doi.org/10.1016/S0893-6080(03)00027-3

Mudunuru, M. K., Son, K., Jiang, P., Hammond, G., & Chen, X. (2022). Scalable deep learning for watershed model calibration. *Frontiers in Earth Science*, *Volume 10 - 2022*. https://doi.org/10.3389/feart.2022.1026479

Thomas, S., Ovchinnikov, M., Yang, F., Voort, D. van der, Cantrell, W., Krueger, S. K., & Shaw, R. A. (2019). Scaling of an atmospheric model to simulate turbulence and cloud microphysics in the pi chamber. *Journal of Advances in Modeling Earth Systems*, *11*(7), 1981–1994.

Wang, A., Jiang, P., Burrows, S., Glienke, S., Ovchinnikov, M., & Mahfouz, N. (2025). *Inverse mapping of the collision kernel and wall flux scaling in a large-scale convection-cloud chamber using local sensors and knowledge-informed deep learning*.

Wang, Aaron, Krueger, S., Chen, S., Ovchinnikov, M., Cantrell, W., & Shaw, R. A. (2024). Glaciation of mixed-phase clouds: Insights from bulk model and bin-microphysics large-eddy simulation informed by laboratory experiment. *Atmospheric Chemistry and Physics*, *24*(18), 10245–10260.

Wang, Aaron, Ovchinnikov, M., Yang, F., Cantrell, W., Yeom, J., & Shaw, R. A. (2024). The dual nature of entrainment-mixing signatures revealed through large-eddy simulations of a convection-cloud chamber. *Journal of the Atmospheric Sciences*, *81*(12), 2017–2039.

Wang, Aaron, Ovchinnikov, M., Yang, F., Schmalfuss, S., & Shaw, R. A. (2024). Designing a convection-cloud chamber for collision-coalescence using large-eddy simulation with bin microphysics. *Journal of Advances in Modeling Earth Systems*, *16*(1), e2023MS003734.

Wang, Aaron, Schmalfuß, S., Chandrakar, K. K., Kia, H. Z., Yang, F., Ovchinnikov, M., Shaw, R. A., & Choi, Y. (2025). An intercomparison of wall fluxes in a turbulent thermal convection chamber: Direct numerical simulations and wall-modeled large-eddy simulations enhanced by machine learning. *Physics of Fluids*, *37*(4).

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.*, *55*(4). https://doi.org/10.1145/3514228