The Journal of Open Source Software

MOTrainer: Distributed Measurement Operator Trainer for Data Assimilation Applications

Ou Ku[1], Fakhereh Alidoost[1], Xu Shan[2], Pranav Chandramouli[1], Sonja Georgievska[1], Meiert W. Grootes[1], and Susan Steele-Dunne[2]¶

1 Netherlands eScience Center, Netherlands **2** Delft University of Technology, Netherlands ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 12 February 2024
Published: unpublished

Summary

Data assimilation (DA) is an essential procedure in Earth and environmental sciences, enabling physical model states to be constrained using observational data.

In the DA process, observations are integrated into the physical model through the application of a Measurement Operator (MO) – a connection model mapping physical model states to observations. Researchers have observed that employing a Machine-Learning (ML) model as a surrogate MO can bypass the limitations associated with using an overly simplified MO (B. A. Forman & Xue, 2017; B. Forman & Reichle, 2014; Xue & Forman, 2015).

Statement of Need

A surrogate MO, as a ML model is trained with the assumption that a single MO applies when mapping physical model states to observations. When dealing with a large spatio-temporal scale, multiple mapping processes may exist, prompting consideration for training separate MOs for distinct spatial and/or temporal partitions of the dataset. As the number of partitions increases, a challenge arises in distributing these training tasks effectively among the partitions.

To address this challenge, we developed a novel approach for distributed training of MOs. We present the open Python library `MOTrainer`, which to the best of our knowledge, is the first Python library catering to researchers requiring training independent MOs across extensive spatio-temporal coverage in a distributed manner. `MOTrainer` leverages Xarray's (Hoyer & Joseph, 2017) support for multi-dimensional datasets to accommodate spatio-temporal features of input/output data of the training tasks. It provides user-friendly functionalities implemented with the Dask (Rocklin, 2015) library, facilitating the partitioning of large spatio-temporal data for independent model training tasks. Additionally, it streamlines the train-test data split based on customized spatio-temporal coordinates. The Jackknife method (McCuen, 1998) is implemented as an external Cross-Validation (CV) method for Deep Neural Network (DNN) training, with support for Dask parallelization. This feature enables the scaling of training tasks across various computational infrastructures.

`MOTrainer` has been employed in a study of vegetation water dynamics (Shan et al., 2022), where it facilitated the mapping of Land-Scape Model (LSM) states to satellite radar observations.

Tutorial

The `MOTrainer` package includes comprehensive usage examples, as well as tutorials for:

1. Converting input data to Xarray Dataset format: Example 1 and Example 2;
2. Training tasks on simpler ML models using `sklearn` and `daskml`: Example Notebook;
3. Training tasks on Deep Neural Networks (DNN) using TensorFlow: Example Notebook.

Acknowledgements

The authors express sincere gratitude to the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) and the Netherlands Space Office for their generous funding of the MOTrainer development through the User Support Programme Space Research (GO) call, grant ALWGO.2018.036. Special thanks to SURF for providing valuable computational resources for MOTrainer testing via the grant EINF-339.

We would also like to thanks Dr. Francesco Nattino, Dr. Yifat Dzigan, Dr. Paco López-Dekker, and Tina Nikaein for the insightful discussions, which are important contributions to this work.

References

Forman, B. A., & Xue, Y. (2017). Machine learning predictions of passive microwave brightness temperature over snow-covered land using the special sensor microwave imager (SSM/i). *Physical Geography*, *38*(2), 176–196. https://doi.org/10.1080/02723646.2016.1236606

Forman, B., & Reichle, R. (2014). Using a support vector machine and a land surface model to estimate large-scale passive microwave brightness temperatures over snow-covered land in north america. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*, 1–11. https://doi.org/10.1109/JSTARS.2014.2325780

Hoyer, S., & Joseph, H. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1). https://doi.org/10.5334/jors.148

McCuen, R. H. (1998). *Hydrologic analysis and design*. Prentice Hall. ISBN: 9780131349582

Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. *SciPy*. https://api.semanticscholar.org/CorpusID:63554230

Shan, X., Steele-Dunne, S., Huber, M., Hahn, S., Wagner, W., Bonan, B., Albergel, C., Calvet, J.-C., Ku, O., & Georgievska, S. (2022). Towards constraining soil and vegetation dynamics in land surface models: Modeling ASCAT backscatter incidence-angle dependence with a deep neural network. *Remote Sensing of Environment*, *279*, 113116. https://doi.org/https://doi.org/10.1016/j.rse.2022.113116

Xue, Y., & Forman, B. A. (2015). Comparison of passive microwave brightness temperature prediction sensitivities over snow-covered land in north america using machine learning algorithms and the advanced microwave scanning radiometer. *Remote Sensing of Environment*, *170*, 153–165. https://doi.org/https://doi.org/10.1016/j.rse.2015.09.009