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: 10.21105/joss.06591

Software

- Review
- Repository
- Archive

Editor: Sarath Menon

Reviewers:

- @abhishektiwari
- @WarmCyan

Submitted: 12 February 2024
Published: 22 August 2024

Summary

Data assimilation (DA) is an essential procedure in Earth and environmental sciences, enabling physical model states to be constrained using observational data. (Albergel et al., 2018; Carrassi et al., 2018; Evensen, 2009; Reichle, 2008)

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, trained as a ML model, is generally considered valid within a specific spatio-temporal range. (Reichle, 2008; Shan et al., 2022; Zhou et al., 2008) 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 (Efron, 1982) is implemented as an external Cross-Validation 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 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

Albergel, C., Munier, S., Bocher, A., Bonan, B., Zheng, Y., Draper, C., Leroux, D. J., & Calvet, J.-C. (2018). LDAS-Monde Sequential Assimilation of Satellite Derived Observations Applied to the Contiguous US: An ERA-5 Driven Reanalysis of the Land Surface Variables. *Remote Sensing*, *10*(10). https://doi.org/10.3390/rs10101627

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535. https://doi.org/10.1002/wcc.535

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans* (Vol. 38). SIAM. https://doi.org/10.1137/1.9781611970319

Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, *29*(3), 83–104. https://doi.org/10.1109/MCS.2009.932223

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

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in Water Resources*, *31*(11), 1411–1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Rocklin, M. (2015). Dask: Parallel Computation with Blocked algorithms and Task Scheduling. *SciPy*. https://doi.org/10.25080/majora-7b98e3ed-013

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/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/10.1016/j.rse.2015.09.009

Zhou, Y., McLaughlin, D., Entekhabi, D., & Ng, G.-H. C. (2008). An ensemble multiscale filter for large nonlinear data assimilation problems. *Monthly Weather Review*, *136*(2), 678–698. https://doi.org/10.1175/2007MWR2064.1