The Journal of Open Source Software

DataAssimilationBenchmarks.jl: a data assimilation research framework.

Colin Grudzien [1,2], **Charlotte Merchant** [1,3], **and Sukhreen Sandhu** [4]

1 CW3E - Scripps Institution of Oceanography, University of California, San Diego, USA **2** Department of Mathematics and Statistics, University of Nevada, Reno, USA **3** Department of Computer Science, Princeton University, USA **4** Department of Computer Science and Engineering, University of Nevada, Reno, USA

DOI: 10.21105/joss.04129

Software

- Review
- Repository
- Archive

Editor: Patrick Diehl

Reviewers:

- @peanutfun
- @tmigot

Submitted: 03 December 2021
Published: 11 November 2022

Summary

Data assimilation (DA) refers to techniques used to combine the data from physics-based, numerical models and real-world observations to produce an estimate for the state of a time-evolving random process and the parameters that govern its evolution (Asch et al., 2016). Owing to their history in numerical weather prediction, full-scale DA systems are designed to operate in an extremely large dimension of model variables and observations, often with sequential-in-time observational data (Carrassi et al., 2018). As a long-studied "big-data" problem, DA has benefited from the fusion of a variety of techniques, including methods from Bayesian inference, dynamical systems, numerical analysis, optimization, control theory, and machine learning. DA techniques are widely used in many areas of geosciences, neurosciences, biology, autonomous vehicle guidance, and various engineering applications requiring dynamic state estimation and control.

The purpose of this package is to provide a research framework for the theoretical development and empirical validation of novel data assimilation techniques. While analytical proofs can be derived for classical methods, such as the Kalman filter in linear-Gaussian dynamics (Jazwinski, 2007), most currently developed DA techniques are designed for estimation in nonlinear, non-Gaussian models where no analytical solution may exist. DA methods, therefore, must be studied with rigorous numerical simulation in standard test-cases to demonstrate the effectiveness and computational performance of novel algorithms. Pursuant to proposing a novel DA method, one should likewise compare the performance of a proposed scheme with other standard methods within the same class of estimators.

This package implements a variety of standard data assimilation algorithms, including some of the widely used performance modifications that are used in practice to tune these estimators. This software framework was written originally to support the development and intercomparison of methods studied in C. Grudzien & Bocquet (2022). Details of the studied ensemble DA schemes, including pseudo-code detailing their implementation and DA experiment benchmark configurations, can be found in the above principal reference. Additional details on numerical integration schemes utilized in this package can be found in the secondary reference (Grudzien C. et al., 2020).

Statement of need

Standard libraries exist for full-scale DA system research and development, e.g., the Data Assimilation Research Testbed (DART) (Anderson et al., 2009), but there are fewer standard options for theoretical research and algorithm development in simple test systems. Many basic research frameworks, furthermore, do not include standard operational techniques developed from classical variational methods, due to the difficulty in constructing tangent linear and adjoint codes (Kalnay et al., 2007). DataAssimilationBenchmarks.jl provides one framework for studying sequential filters and smoothers that are commonly used in online, geoscientific prediction settings, including both ensemble methods and variational schemes, with hybrid methods planned for future development.

Comparison with similar projects

Similar projects to DataAssimilationBenchmarks.jl include the DAPPER Python library (Raanes & others, 2022), DataAssim.jl used by Vetra-Carvalho et al. (2018), and EnsembleKalmanProcesses.jl of the Climate Modeling Alliance (Dunbar & others, 2022). These alternatives are differentiated primarily in that:

- DAPPER is a Python-based library which is well-established, and includes many of the same estimators and models. However, numerical simulations in Python run notably slower than simulations in Julia when numerical routines cannot be vectorized in Numpy (Bezanson et al., 2017). Particularly, this can make the wide hyper-parameter search intended above computationally challenging without utilizing additional packages such as Numba (Lam et al., 2015) for code acceleration.
- DataAssim.jl is another established Julia library, but notably lacks an implementation of variational and ensemble-variational techniques.
- EnsembleKalmanProcesses.jl is another established Julia library, but specifically lacks traditional geoscientific DA approaches such as 3D-VAR and the ETKF/S.

Future development

The future development of the DataAssimilationBenchmarks.jl package is intended to expand upon the existing, variational and ensemble-variational filters and sequential smoothers for robust intercomparison of novel schemes. Additional process models and observation models for the DA system are also in development.

Acknowledgements

Colin Grudzien developed the numerical code for the package's Julia type optimizations for numerical schemes and automatic differentiation of code, the ensemble-based estimation schemes, the observation models, the Lorenz-96 model, the IEEE 39 Bus test case model, and the numerical integration schemes for ordinary and stochastic differential equations. Charlotte Merchant developed the numerical code for implementing variational data assimilation in the Lorenz-96 model and related experiments. Sukhreen Sandhu supported the development of the package structure and organization. All authors supported the development of the package by implementing test cases and writing software documentation. This work was supported by the University of Nevada, Reno, Office of Undergraduate Research's Pack Research Experience Program which supported Sukhreen Sandhu as a research assistant. This work was supported by the Center for Western Weather and Water Extremes internship program which supported Charlotte Merchant as a research assistant. This work benefited from the DAPPER library which was referenced at times for the development of DA schemes. The authors would like to thank the two handling editors Bita Hasheminezhad and Patrick Diehl, and the two named referees Lukas Riedel and Tangi Migot, for their comments, suggestions, and valuable advice which strongly improved the quality of the paper and the software.

References

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, *90*(9), 1283–1296. https://doi.org/10.1175/2009BAMS2618.1

Asch, M., Bocquet, M., & Nodet, M. (2016). *Data assimilation: Methods, algorithms, and applications*. SIAM. https://doi.org/10.1137/1.9781611974546

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

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

Dunbar, O., & others. (2022). *CliMA/EnsembleKalmanProcesses.jl: v0.10.0* (Version v0.10.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.7036069

Grudzien, C., & Bocquet, M. (2022). A fast, single-iteration ensemble Kalman smoother for sequential data assimilation. *Geoscientific Model Development*, *15*(20), 7641–7681. https://doi.org/10.5194/gmd-15-7641-2022

Grudzien, C., Bocquet, M., & Carrassi, A. (2020). On the numerical integration of the Lorenz-96 model, with scalar additive noise, for benchmark twin experiments. *Geoscientific Model Development*, *13*(4), 1903–1924. https://doi.org/10.5194/gmd-13-1903-2020

Jazwinski, A. H. (2007). *Stochastic Processes and Filtering Theory*. Dover Publications, Incorporated. ISBN: 9780486785349

Kalnay, E., Li, H., Miyoshi, T., Yang, S. C., & Ballabrera-Poy, J. (2007). 4-D-Var or ensemble Kalman filter? *Tellus A: Dynamic Meteorology and Oceanography*, *59*(5), 758–773. https://doi.org/10.1111/j.1600-0870.2007.00261.x

Lam, S. K., Pitrou, A., & Seibert, S. (2015). Numba: A LLVM-based Python JIT compiler. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 1–6. https://doi.org/10.1145/2833157.2833162

Raanes, P. N., & others. (2022). *Nansencenter/DAPPER: Version 1.2.1,*. Zenodo. https://github.com/nansencenter/DAPPER

Vetra-Carvalho, S., Van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, M. U., Brasseur, P., Kirchgessner, P., & Beckers, J. M. (2018). State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. *Tellus A: Dynamic Meteorology and Oceanography*, *70*(1), 1–43. https://doi.org/10.1080/16000870.2018.1445364