Data-Driven Reduced-Order Model for Atmospheric CO_2 Dispersion

Pedro R. B. Rocha,[1,2] **Marcos S. P. Gomes,**[2] **João L. S. Almeida,**[1]
Allan M. Carvalho,[1] **Alberto C. Nogueira Jr.**[1]

[1]IBM Research Brazil,
[2]Pontifical Catholic University of Rio de Janeiro
pedro.rocha@ibm.com, mspgomes@puc-rio.br, joao.lucas.sousa.almeida@ibm.com,
allancarvalho@ibm.com, albercn@br.ibm.com

Abstract

Machine learning frameworks have emerged as powerful tools for the enhancement of computational fluid dynamics simulations and the construction of reduced-order models (ROMs). The latter are particularly desired when their full-order counterparts portray multiple spatiotemporal features and demand high processing power and storage capacity, such as climate models. In this work, a ROM for CO_2 dispersion across Earth's atmosphere was built from NASA's gridded daily OCO-2 carbon dioxide assimilated dataset. For that, a proper orthogonal decomposition was performed, followed by a non-intrusive operator inference (OpInf). This scientific machine learning technique was capable of accurately representing and predicting the detailed CO_2 concentration field for about one year ahead, with a normalized root-mean-square error below 5%. It suggests OpInf-based ROMs may be a reliable alternative for fast response climate-related predictions.

Introduction

Physics-informed machine learning (PIML) algorithms, which blend data-driven modeling with information about the physics, have been widely employed for improving spatiotemporal forecasts in the Earth and climate sciences (Reichstein et al. 2019; Schneider, Jeevanjee, and Socolow 2021; Cortés-Andrés et al. 2022; Willard et al. 2022; Kashinath et al. 2021). Since purely data-driven machine learning (ML) models are not often generalizable beyond the training interval due to their misspecification or data distribution shifts, violating fundamental principles, additional constraints are required. By incorporating physical laws on the ML model, it has a considerable gain in performance and robustness, becoming more reliable, generalizable and explainable.

One of the primary goals within the PIML field is to build reduced-order models (ROMs) for dynamical systems, which are more computationally efficient than their full-order counterparts in spite of possibly being less precise (Willard et al. 2022). In general, these models are suitable in control, optimization and uncertainty quantification problems, where multiple high-fidelity numerical simulations are needed. Regarding the climate sciences, building ROMs becomes notably challenging due to the Earth system's broad range of scales in space and time and to computational limitations imposed by its very high number of degrees of freedom. It is worth mentioning that, when considering state variables defined mainly by small-scale processes, it is of paramount importance to downscale them to a finer resolution before proceeding to the dimensionality reduction, when many features of the system are inevitably lost.

In the last decade, it was developed a data-driven reduced-order modeling framework based on a non-intrusive operator inference (OpInf) (Peherstorfer and Willcox 2016). This method postulates the shape for the ROM operators based on the knowledge that most physical equations, including those related to fluid flows, are second-order nonlinear. Besides, since OpInf only relies on spatiotemporal data provided by high-fidelity simulations or experimental measurements, it is quite straightforward to be employed in a myriad of dynamical systems to leverage new scientific discoveries. In addition to that, it is computationally more efficient and robust than conventional deep learning techniques, such as echo-state networks (Nogueira Jr et al. 2021), since it requires a much lower number of hyperparameters to be tuned and has a better extrapolation capability. Recently, it was successfully applied in a complex multiscale combustion problem (Swischuk et al. 2020; McQuarrie, Huang, and Willcox 2021).

In this work, to further explore the potential of the OpInf in the field of climate sciences, the gridded daily OCO-2 carbon dioxide assimilated dataset (Weir and Ott 2022), which contains the CO_2 concentration field around the Earth, was considered. The main goal was to obtain a reliable OpInf-based ROM for this field and compare it with the original dataset. To reduce the dimensionality of the system, a proper orthogonal decomposition (POD) was performed. This technique originates from turbulence studies and has been extensively explored in the literature (Lumley 1967).

Data Reduction

The main idea behind POD consists in decomposing a given spatiotemporal vector field $\mathbf{u}(\mathbf{x}, t)$ into a set of spatial functions $\boldsymbol{\phi}_k(\mathbf{x})$, or POD modes, and their respective time coefficients $\alpha_k(t)$ (Weiss 2019). Then, the vector field $\mathbf{u}(\mathbf{x}, t)$ is written as

$$\mathbf{u}(\mathbf{x},t) = \sum_{k=1}^{\infty} \alpha_k(t)\boldsymbol{\phi}_k(\mathbf{x}), \tag{1}$$

where vector (or matrix) quantities are represented in bold.

Although this decomposition may be carried out in different ways, all of them must respect two basic conditions: the spatial functions have to be orthonormal and the first r POD modes must capture the highest possible system's energy, where r is an arbitrary integer. Here, the POD was performed via the principal component analysis (PCA) from Scikit-learn (Pedregosa et al. 2011), which uses the singular value decomposition (SVD) of the data to represent them in a lower-dimensional space (a.k.a. latent space). It should be highlighted that, before applying PCA to the data, they were normalized in the interval $[-1, 1]$ and then centered.

The POD basis V_r, onto which the CO_2 concentration training dataset is projected, is comprised by the r dominant POD modes. In other words, if $C \in \mathbb{R}^{n_t \times n_x}$ is the dataset matrix that contains the spatiotemporal concentration field, then the latent field variables $\widehat{q}(t)$, along the training interval, are obtained by multiplying C by V_r. Here, n_t is the number of timesteps and n_x is the number of grid points. The OpInf technique, to be described next, is then applied to the latent field variables. Note that, before applying PCA to the data, they were normalized from -1 to 1 and then centered.

Non-Intrusive Operator Inference (OpInf)

The ROM for the atmospheric CO_2 dispersion around the Earth was constructed through the OpInf approach (Peherstorfer and Willcox 2016). Since this physical phenomenon is mainly governed by advection and diffusion processes that move the carbon dioxide gas from one place to another, higher-order nonlinearities (cubic and above) were neglected and no forcing term was considered. Then, the general form of the OpInf-based ROM is written as

$$\frac{d}{dt}\widehat{\boldsymbol{q}}(t) = \widehat{\boldsymbol{c}} + \widehat{\boldsymbol{A}}\widehat{\boldsymbol{q}}(t) + \widehat{\boldsymbol{H}}(\widehat{\boldsymbol{q}}(t) \otimes \widehat{\boldsymbol{q}}(t)), \tag{2}$$

where $t \in [t_0, t_f]$, t_0 and t_f are the initial and final time instants, the initial state of the system $\widehat{\boldsymbol{q}}(t_0)$ is known, $\widehat{\boldsymbol{c}} \in \mathbb{R}^r$, $\widehat{\boldsymbol{A}} \in \mathbb{R}^{r\times r}$ and $\widehat{\boldsymbol{H}} \in \mathbb{R}^{r\times r(r+1)/2}$ are the operators to be inferred and the symbol $\otimes$ refers to the Kronecker product.

With the normalized latent field variables along the training interval and their numerically computed time derivatives, it was possible to find the operators $\widehat{\boldsymbol{c}}$, $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{H}}$ by solving a data-driven least-squares regression problem with regularization to avoid overfitting. Then, with optimal regularizers and operators and given the state of the reduced system at $t = t_0$, it was possible to integrate Eq. (2) and obtain all the latent field variables from t_0 to t_f. Finally, these variables were projected back onto the original space, i.e., they were multiplied by V_r^T. To assess the predictive capabilities of the constructed ROM, normalized root-mean-square errors, ϵ_{ROM}, were calculated across the spatiotemporal domain by

$$\epsilon_{ROM} = \frac{1}{\Delta c}\sqrt{\frac{\sum_{j=1}^{n_x}\sum_{i=1}^{n_t}(c_{i,j}^{ROM} - c_{i,j})^2}{n_x n_t}}, \tag{3}$$

where $\Delta c = max(c_{i,j}) - min(c_{i,j})$, $c_{i,j}$ are the elements of the original matrix C, while $c_{i,j}^{ROM}$ corresponds to the concentration field computed by the ROM.

OCO-2 Carbon Dioxide Assimilated Dataset

This dataset from NASA provides a high-quality estimation for the atmospheric CO_2 concentration around the Earth (Weir and Ott 2022). It combines space-based measurements with state-of-the-art data assimilation techniques to handle instruments' inability to see through clouds and thick aerosols. The data cover the period going from 2015.01.01 to 2021.10.30 in a daily basis, totalizing $2,495$ timesteps (or snapshots). The spatial resolution is of $0.5°$ along the Earth's latitudinal axis and of $0.625°$ along the longitudinal one. Then, there is a total of $361 \times 576 = 207,936$ grid points containing the CO_2 data.

Results and Discussion

From $2,495$ snapshots, $2,000$ were used to train the PIML algorithm. To build a data-driven reduced-order model for the atmospheric CO_2 dispersion around the Earth, the training data was initially reduced from $207,936$ dimensions to their five most relevant ones, keeping 98.9% of the accumulated modal energy. This great value is justified by the fact that the average CO_2 concentration highly dominates over its seasonal and spatial variations. Besides, the most dominant mode alone portrays 87.5% of the system's modal energy.

After the dimensionality reduction, a least-squares regression with regularization is applied, as discussed previously. Figure 1 shows that the field variables in latent space obtained from the inferred operators were well approximated by the OpInf-based ROM along the training interval. Both the first and the fourth reduced variables are exhibited in this figure. Note that the ROM neglected the approximation of the noisy pattern for the fourth variable, a consequence of the POD reduction. Also, the model has difficulty in accurately predicting peaks and valleys of the high amplitude oscillations, as observed for the same variable. Such smoothing behavior is commonly seen in ML models.

The original CO_2 concentration field was reconstructed from the latent variables and then compared against the original observation. Figure 2 shows this comparison for a 338-days-ahead forecast. Visually, it may be seen that the ROM captures the main features of the dispersion. The normalized root-mean-square error for the testing interval, computed through Eq. (2), was about 4%. This small value is an indicative that the model is quite robust.

Conclusion

The capabilities of the data-driven reduced-order model based on a non-intrusive operator inference approach for

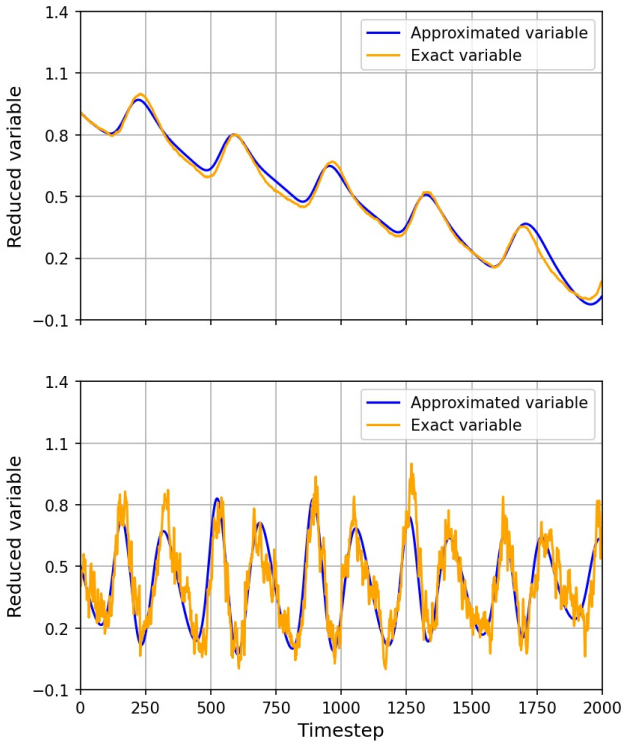

Figure 1: Approximated and exact reduced variables along the training interval. On the top, the most relevant reduced variable, which carries 87.5% of the system's modal energy. On the bottom, the fourth one

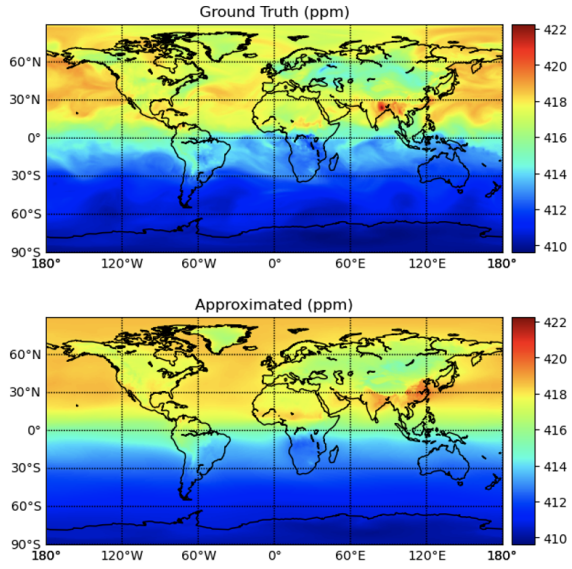

Figure 2: CO_2 concentration field around the Earth according to the high-fidelity data (top) and to the reduced-order model (bottom) for the 338^{th} testing snapshot

the atmospheric CO_2 dispersion around the Earth was assessed. It presented excellent predictive capabilities for this physical phenomenon in addition to being quickly deployed, with normalized root-mean-square errors below 5% for the testing interval. This physics-informed machine learning method seems to be adequate for large-scale climate systems mainly governed by advection and diffusion processes. In practical terms, the OpInf-based ROM is well suited for uncertainty quantification of climate-related predictions.

Acknowledgments

The authors would like to thank IBM and the Brazilian agencies CAPES and CNPq for the financial support to this work.

References

Cortés-Andrés, J.; Camps-Valls, G.; Sippel, S.; Sz'ekely, E.; Sejdinovic, D.; Diaz, E.; P'erez-Suay, A.; Li, Z.; Mahecha, M.; and Reichstein, M. 2022. Physics-aware nonparametric regression models for Earth data analysis. *Environmental Research Letters*, 17(5): 054034.

Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.-L.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A.; Manepalli, A.; Chirila, D.; Yu, R.; Walters, R.; White, B.; Xiao, H.; Tchelepi, H.; Marcus, P.; Anandkumar, A.; and Prabhat, M. 2021. Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379: 20200093.

Lumley, J. L. 1967. The structure of inhomogeneous turbulent flows. *Atmospheric Turbulence and Radio Wave Propagation*.

McQuarrie, S. A.; Huang, C.; and Willcox, K. E. 2021. Data-driven reduced-order models via regularised Operator Inference for a single-injector combustion process. *Journal of the Royal Society of New Zealand*, 51(2): 194–211.

Nogueira Jr, A. C.; Carvalho, F. C.; Almeida, J. L. S.; Codas, A.; Bentivegna, E.; and Watson, C. D. 2021. Reservoir Computing in Reduced Order Modeling for Chaotic Dynamical Systems. In *International Conference on High Performance Computing*, 56–72. Springer, Cham.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Édouard Duchesnay. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12(85): 2825–2830. In press.

Peherstorfer, B.; and Willcox, K. 2016. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306: 196–215.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*.

Schneider, T.; Jeevanjee, N.; and Socolow, R. 2021. Accelerating progress in climate science. 74: 44–51.

Swischuk, R.; Krämer, B.; Huang, C.; and Willcox, K. 2020. Learning physics-based reduced-order models for a single-injector combustion process. *AIAA Journal*, 58: 1–15. In press.

Weir, B.; and Ott, L. 2022. OCO-2 GEOS Level 3 daily, 0.5x0.625 assimilated CO2 V10r. https://disc.gsfc.nasa.gov/datasets/OCO2_GEOS_L3CO2_DAY_10r/summary. Accessed: 2022-07-19.

Weiss, J. 2019. A Tutorial on the Proper Orthogonal Decomposition. AIAA 2019–3333. American Institute of Aeronautics and Astronautics. Available Open Access acceptedVersion at https://depositonce.tu-berlin.de/handle/11303/9456.

Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; and Kumar, V. 2022. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Computing Surveys*.