https://esd.copernicus.org/preprints/esd-2024-30/esd-2024-30.pdf

The paper "***Potential for Equation Discovery with AI in the Climate Sciences***" is a vital discussion topic for advancing climate research. It's clear that there are infinitely many more non-linear formulations than the linear set of possibilities that humans are comfortable with solving. Fluid dynamics *a la* Navier-Stokes by itself contains many non-linear elements that have not been completely explored due to a lack of ability to solve in a closed form. The paper suggests an important possible constraint to apply: *"For physical systems involving fluid flows where the underlying equations are known to be energy-preserving, although also nonlinear".*

And that's where artificial neural networks and symbolic regression (i.e. equation discovery) comes into play. There are really few other alternatives outside of tedious human trial & error that are available to both (1) fully explore the combinatorial solution space and (2) incorporate numerical solvers to train the possible solutions to fit the available data using appropriate metrics for plausibility and precision.

The paper as is falls short on two fronts, one of which the authors' themselves highlight. The first can be remedied by citing the importance of *cross-validation* (CV) strategies. The success of machine learning is in part due to how CV can separate the wheat from the chaff in potential solutions. Yet, nowhere in the text is cross-validation mentioned, and this is a vital part of equation discovery, as an optimal CV algorithm+metric is necessary to isolate candidate solutions along a Pareto front of complexity (1/plausibility) vs precision. Neural networks can fit just about any curve, so CV approaches to equation discovery help to eliminate those that are the result of over-fitting. Suggest Ref [1] as a citation starting point.

The second front is based on the authors' statement *"It is relatively easy to set aspirations for implementing AI methods in climate science, rather than performing the analysis itself"*. I read this as a call to just do it instead of dreaming it, or as the thespian philosopher Christopher Walken said: "If you want to learn how to **build a house**, then **build a house**. Don't ask anybody. Just **build a house**." The paper suggested
"We discuss the potential application of AI-led equation discovery to three Earth system components. In each example, there is presently a deficiency in understanding, causing uncertainty in the representation of processes by equations. Each application falls into one of three categories. "

Instead, I would recommend three Earth system components to evaluate: solid body, atmosphere (gas fluid), and ocean (liquid fluid). In our text Mathematical Geoenergy, P. Pukite, D. Coyne, D. Challou (Wiley/AGU, 2019), we describe novel equation-based models for the Earth's Chandler wobble (solid body), QBO (atmosphere),and ENSO (ocean). The original nonlinear models were derived from simplifying Euler equations of motion for the Chandler wobble, and Laplace's Tidal Equations, which are simplified Navier-Stokes, for QBO and ENSO. We attain excellent agreement against observations in each case, and this extends to other climate indices such as AMO and PDO. See Figures 1..X at the end of this review.

Over the past few years, I have tried various machine learning approaches including neural networks and symbolic regression to observe if they would "discover" the same equation solutions I had formulated and applied. First, it's clear that neural networks can't do the job as they train only on their own data-set as supplied, and so won't automatically pull in all the tidal time-series data available. This is the *closed-world assumption* (CWA) problem well-known in AI circles for years, see Ref [2]. Neural networks will fit the data, but it's all based on dreaming up patterns from the data instead of tracing it back to a non-linear modulation from an external forcing. Alas, that external data set doesn't exist in the training data, so it gets ignored.

The symbolic regression/equation discovery approaches do an arguably better job. Although they also suffer from the CWA problem, they can make up for it by creating symbolic expressions from their library of primitive mathematical operators to draw from, such as creating a tidal forcing from (1) the time base, (2) arbitrary constants, and (3) sinusoidal primitives such as sin() and cos(). So, in terms of results, the frequencies from tidal factors do emerge in a symbolic regression fit to QBO, yet they are not directly harmonically-related due to the intrinsic non-linearity of the equation solutions! Thus, they may easily get overlooked when the symbolic regression results are deconstructed, as it also requires knowledge of nonlinear signal processing concepts such as aliasing and side-banding. That's what I have found straightforwardly in the Chandler wobble and QBO results, and with more of a challenge in the oceanic indices such as ENSO. The symbolic regression tools that I have evaluated include Eureqa, PySR, and TuringBot, Ref [3].

And this reflects back on the importance of cross-validation approaches and the selection of correlation metrics, including those that have proved valuable in machine learning in the context of noise and uncertainty, such as dynamic time warping - Ref [4] and complexity-invariance distance - Ref [5]. The results of symbolic regression depend on the best metric for the data, as some may prove too stiff to emerge from a local optima.

I agree with the paper that the focus on statistical machine learning to model climate variation is misguided, as it is more evident that large scale behaviors that are the result of collective deterministic actions describe better the standing wave models of ENSO and QBO. These will show the detail and variety in waveforms captured by wave equations, not the smeared responses captured by statistical ensembles.

Moreover (and finally), it is difficult to get a new paradigm accepted in geophysics fields such as climate science unless the results are beyond reproach. The complete lack of controlled experiments to test novel equation-based models means that claims of excellent agreement are dealt with suspicion. It is costly in terms of money and time to wait years for predictive models to come true, so the hope is that cross-validation results can conclusively demonstrate a new equation formulation has merit.

Ref

[1] Sweet, L., C. Müller, M. Anand, and J. Zscheischler, 2023: Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. *Artif. Intell. Earth Syst.*, **2**, e230026, https://doi.org/10.1175/AIES-D-23-0026.1.
[2] Zhu, Fei, Shijie Ma, Zhen Cheng, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. "Open-world machine learning: A review and new outlooks." *arXiv preprint arXiv:2403.01759* (2024).
[3] In my opinion, Eureqa did the best job but it has not been available for use since 2017 as it was sold to an AI firm for proprietary use. https://en.wikipedia.org/wiki/Eureqa , https://turingbotsoftware.com/ which is a attempted clone of Eureqa. https://github.com/MilesCranmer/PySR
[4] Li, Hailin. "Time works well: Dynamic time warping based on time weighting for time series data mining." *Information Sciences* 547 (2021): 592-608.
[5] Batista, Gustavo EAPA, et al. "CID: an efficient complexity-invariant distance for time series." *Data Mining and Knowledge Discovery* 28 (2014): 634-669.

Models of geophysical behaviors

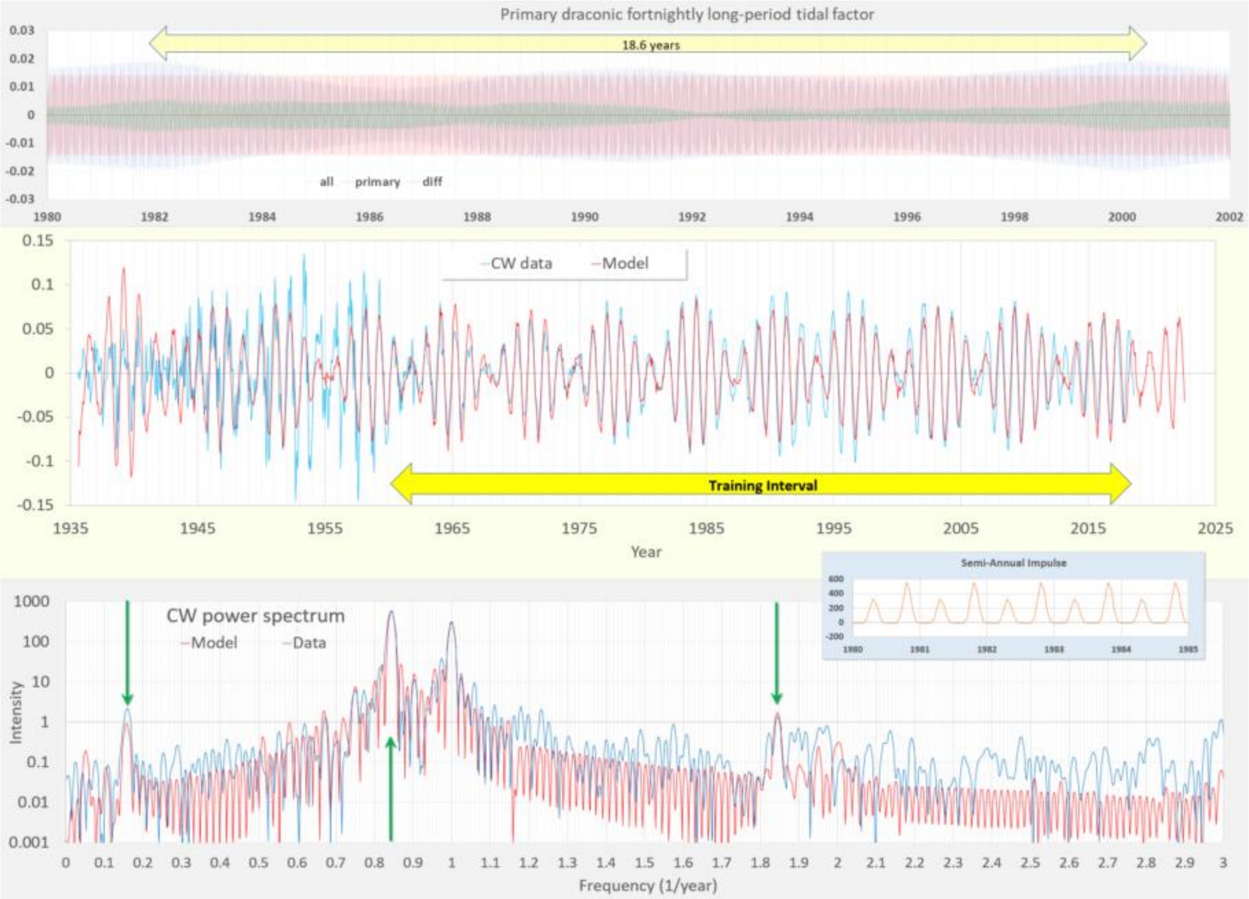

Figure 1 : Cross-validated **Chandler wobble** model of luni-solar torqued Euler equations

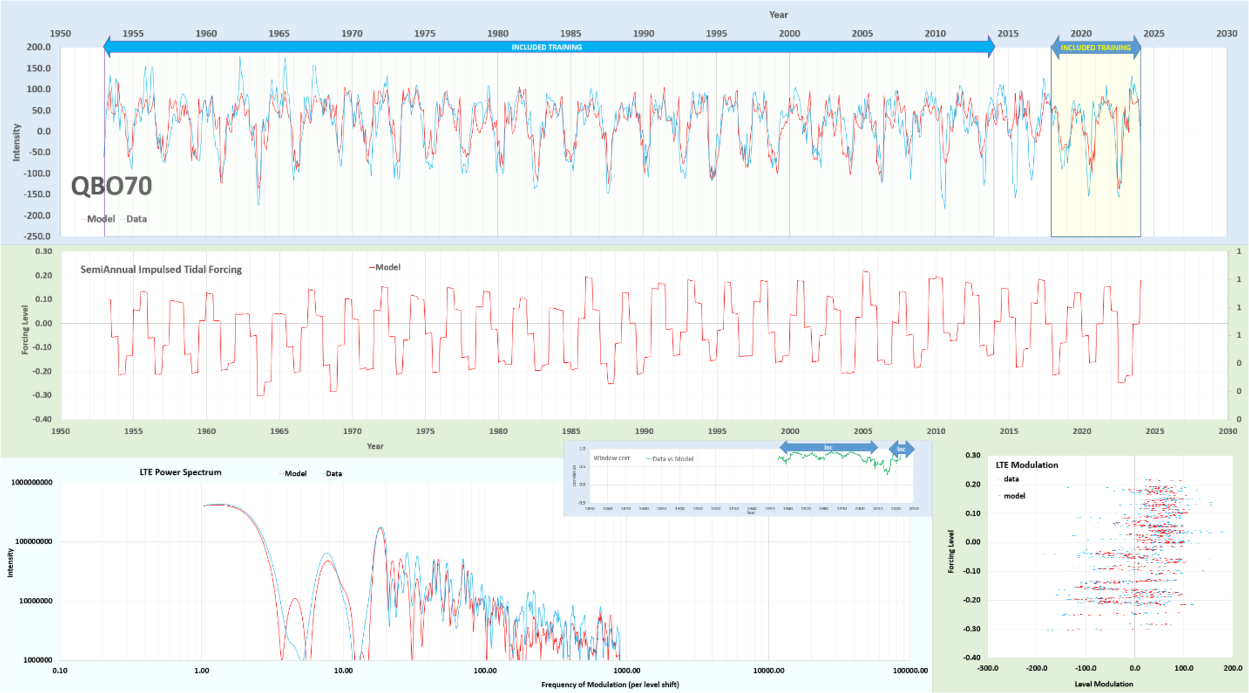

Figure 2: Cross-validated **QBO** model at 70hPa of luni-solar forced Laplace’s Tidal Equations

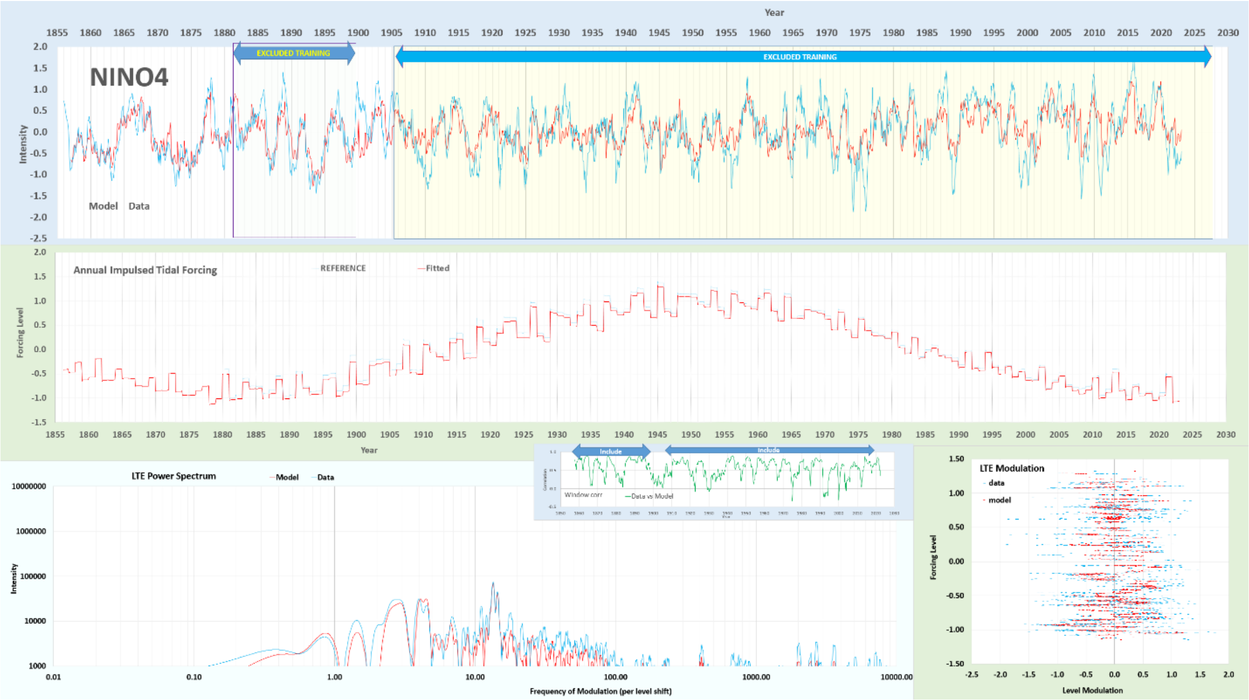

Figure 3: Cross-validated **ENSO** NINO4 modelof luni-solar forced Laplace’s Tidal Equations

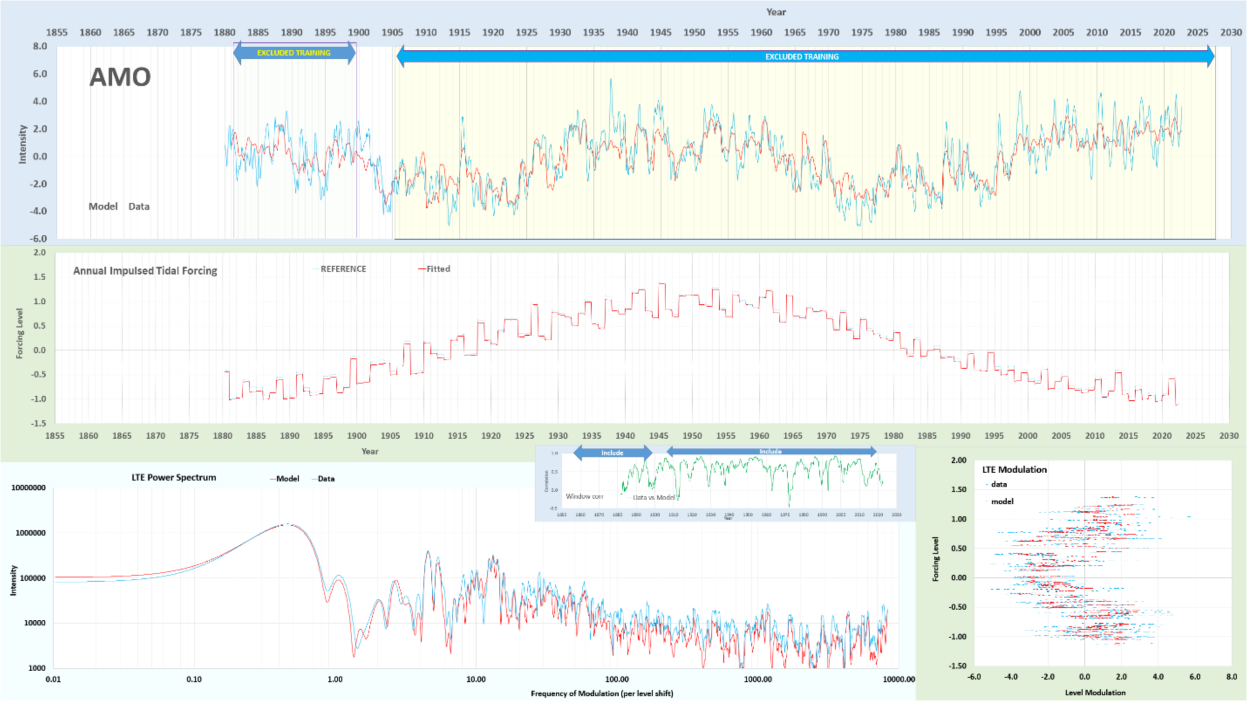

Figure 4 : Cross-validated **AMO** model of luni-solar forced Laplace's Tidal Equations

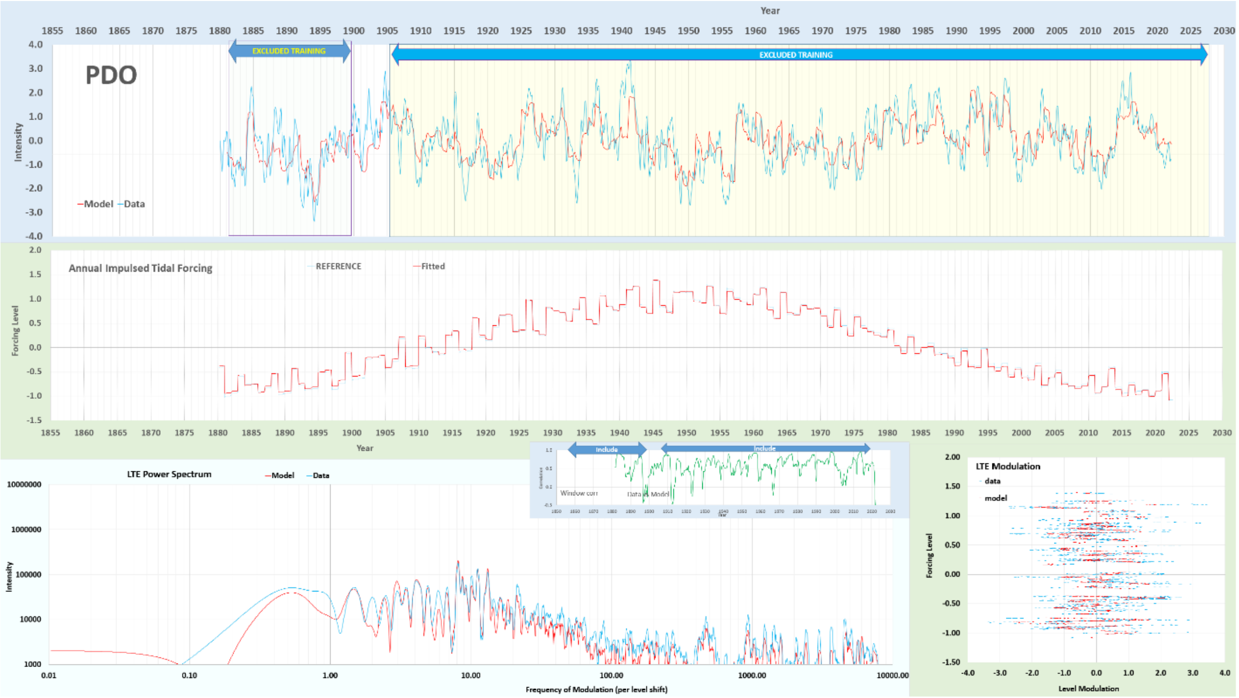

Figure 5 : Cross-validated **PDO** model of luni-solar forced Laplace's Tidal Equations