Automated Discovery of DOminaNt physics Informed Surrogates (ADDONIS) Framework for Improving Water Cycling Predictability

Kenneth (Chad) Sockwell, **SNL**, 1442: Computational Mathematics
Pavel Bochev, **SNL**, 1400: Center for Computing Sciences
J. Derek Tucker, **SNL**, 6673: Statistical Sciences; Department of Statistics, University of Illinois
Paul Kuberry, **SNL**, 1442: Computational Mathematics

Focal Areas

The primary alignment of the proposed framework is with focal area 2, because it guides the construction of a hierarchy of predictive models generated from Statistical and AI/ ML techniques. There are connections to focal areas 1 and 3 because surrogates allow for coupled data assimilation techniques where the full model is too expensive (1) and the discovered balance laws can be used as an insightful analysis tool to validate the structure of models (3).

Science Challenge

Can the predictability of the water cycle be improved through the use of data-driven discovered balance-law surrogates, comprised of dominant physics existing within coupled models, allowing vast uncertainty quantification (UQ), data-assimilation, calibration / parameter estimation, and validation that is far too costly for a fully coupled earth system model (ESM)? The design of situationally informed and highly efficient, lower-fidelity models serving as a surrogate to the coupled ESM provides a feasible way to generate large ensembles to investigate extremes in the statistics of the model. Additionally, the surrogate models can be constructed from both simulation and experimental data in an automatic way, leading to a model-experiment coupling (MODEX) like framework. The MODEX paradigm aims to improve predictive understanding through a workflow which interweaves experiment and simulation at all steps possible.

Rationale

Increasing predictability, decreasing uncertainty, and investigating extremes in the integrated water cycle requires many queries to the fully coupled ESM. Examples include data assimilation, uncertainty quantification (UQ) and optimization techniques, as well as generating large amount of training data for neural networks. All of these cases require large ensembles of data. Currently, running even small ensembles for the DOE's High-Resolution E3SM climate model is not feasible on today's HPCs. It is not expected that large ensembles will be enabled with future computing systems in the next ten years. Highly efficient, lower-fidelity models with correct physics provide a way to resolve this performance bottleneck. It might be the case that a small set of physics or simplified balance laws within the coupled model completely describe some process or physical regime, such that improved efficiency can be realized by ignoring other less important processes. Balance laws are described in box **B1**. Typically, asymptotic analysis is relied upon to discover simplified balance laws. Asymptotic analysis provides a reduction of physics, beyond that in conventional reduced modeling strategies, through the identification of the dominant physics. This technique can be used to **discover simplified yet highly efficient driver-response relations represented by simplified balance laws** within the coupled model. Simplified Balance laws within the ESM can be used to inform the construction of highly efficient, lower-fidelity surrogate models to represent the ESM dynamics. The surrogate models will preserve the most dominant physics relevant for a specific science question. However, **performing asymptotic analysis using traditional techniques is not feasible for the entire coupled ESM** with all physics and parameterizations included. E3SM would benefit from a **data-based analog to asymptotic analysis** which

B1: The idea of dominant balance can be exemplified with the Navier-Stokes equations for a velocity field,

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p + \nu \Delta u = 0$$

For instance, it is well known that the limit of growing viscosity yields the linear Stokes equations, and the Euler equations in the limit of zero viscosity. The terms which remain present in the asymptotic limit balance to zero and describe the dominant balance in the respective physical regime.

DOI: https://doi.org/10.2172/1769678
DOE BER, AI4ESP 2021 #AI4ESP1124 (https://www.ai4esp.org)

can be used to construct surrogates. A hierarchy of data-driven / discovered models naturally complements any ESM model and naturally leads to a validation framework where the theoretical and observation-based model can be compared to evaluate biases. Moreover, any AI/ML generated hierarchy of balance-law surrogates can be used in combination with the full models such as E3SM or any of its individual components, only requiring coupling technology from traditional to data driven models. The assembled team has a diverse and interdisciplinary background consisting of ML [1] [2], statistics [3], geometrical and functional data analysis, climate models [4], reduced order modeling [5], modeling coupling [6] [7], and rigorous mathematical analysis. All of these skills will be required to design and develop the proposed framework.

A concrete example of the framework can be given for ensemble generation in the interest of extremes in the water cycle. Suppose one is interested in precipitation extremes over the east coast of the United States over decades. The framework analyzes coupled ESM simulation or observational data of the larger region to discover dominant physics through balance laws. Simplified balance laws could potentially be identified between the ocean-air-land system over different local regions or in time. For instance, in [8], a data-driven technique was used to identify local areas of geostrophic balance from 1/25° HYCOM reanalysis data. The authors also perform this analysis and recover classical asymptotic results for other systems. Local balance laws over different regions could then be reduced into surrogate models. The surrogate models then provide a feasible way to construct large ensembles, that encapsulate dominant coupled physics, used to investigate extremes.

These situational and domain physics aware surrogates can be used in tandem with E3SM to advance science campaigns by providing a hierarchy of automatically generated models to increase prediction skill and lower uncertainty. The ADDONIS framework is strongly aligned with the DOE's EESSD ***Data-Model Integration Scientific Grand Challenge***, directly aimed at ***Associate Research Goals*** 2-4 (**ARG**) and requires goal 1 as a supporting technology, within the EESSD strategic plan: (**ARG 2**) The data-driven surrogates can provide validation and UQ for the model, (**ARG 3**) directly informing the model and guiding further experiments to further reduce uncertainty and (**ARG 4**) generating a hieratical of suite of surrogates, designed to be accurate for different scales, physics, or science questions. The ability to construct driver-response surrogates representing coupled models and providing analysis of dominant physics has the potential to **strongly impact all other *Grand Challenges***.

Narrative

Physical models are derived from first principles such as balance laws, which often assume the form of conservation laws. It is typically the case that some processes dominate or strongly contribute to the model in certain physical regimes or for certain parameter choices. This is realized within the mathematical model by the balancing (summing to zero) of a subset of terms in the homogenous from of the mathematical model. Using the principle of parsimony, these simplified balance laws are used where appropriate fidelity is upheld. Within the climate system, there are many balanced flows that can be derived from asymptotic analysis, such as geostrophic balance and corresponding quasi-geostrophic models. Asymptotic analysis has been a powerful tool for model development for centuries. However, these classical techniques become very difficult and require heuristics and/or assumptions with limited physical validity when highly coupled multi-physics systems are considered, such as the fully coupled ESM. Furthermore, classical techniques are not applicable to experimental data or observations to develop models. The following question is the central theme of this paper: **Are there dominant balance laws, existing within the coupled climate system and spanning various components, that can be learned automatically and then utilized to construct a hierarchical suite of highly efficient lower-fidelity surrogate models built for specific physical regimes or science questions**? In [8], the authors describe a technique to learn, in an unsupervised manner, the dominant physics in various spatio-temporal regimes for different selections of parameters. The technique that Callahan provides enables an additional level of reduction of the physics, beyond the parameter-space or time-dimension reduction seen in typical reduced-

order-modeling (ROM). Moreover, the important physics and less important (to be neglected) physics can be identified in different spatial and temporal regions. A framework, potentially in-situ with simulations and observation sensors, could lead to a MODEX like paradigm that automatically generates a hierarchy of surrogate models for applications like data-assimilation or UQ, providing verification of the ESM.
Here we coin the ADDONIS framework, which provides a novel strategy to integrate observational and simulation data to automatically synthesize coupled surrogate models by discovering balance laws and dominant physical modes with statistical and ML methodologies on multiscale spatio-temporal data. In essence, the ADDONIS framework learns dominant physics and what physics can be neglected and uses this information to automatically construct surrogates. The discovery of balance laws and dominant physics for different physical regimes (various model and parameter choices) and scales naturally enables the validation and calibration of computational models through comparison.

The ADDONIS framework, which uses Callahan et al. 2020 as a starting point, can be compactly described in the following **four steps**, each requiring its own technical advancements in a synergistic manner to facilitate the framework's success:

1. **Data Acquisition**
2. **Equation Space Choice**
3. **Clustering and Dimensionality Reduction**
4. **Building Models or Analysis**

The first step, **Data Acquisition**, consists of obtaining either observation or simulation spatio-temporal data. Applications that allow for the use of simulation data are more straightforward because the output data can be tailored to the framework nor are there large gaps in the data from sensor limitations. Of course, using observation data is more desirable in the framework and certain challenges accompanying the use of observation data must be overcome. Recently, and continuing into the future, the DOE's Earth System observation capabilities have largely increased with facilities like ARM and big data sets for ocean and atmosphere currents, temperatures, and chemistry composition, etc. are available. As these capabilities increase in the future and better sensors are developed, deficiencies of observation data such as insufficient number of sensors or high uncertainty in the measurements are lessened. The second step, **Equation Space Choice**, defines the ansatz, or assumed form, the dominant balance law will be in. A typically choice for the ansatz is the highest fidelity model possible, such as the Navier-Stokes equations in the case of fluids, shown in box **B1**. For climate models, the ESM itself can serve as the chosen ansatz, as it is expected that some of the complex processes can be neglected in specific situations. Moreover, if no high-fidelity model is known the data can be regressed or mapped into processes which are expected to balance a-priori using system identification methods. We conjecture that the dynamics in local regions can be described by a few terms in equations. The third step utilizes data-driven techniques, which discover dominant physics by identifying clusters in the **equation space** consisting of a basis corresponding to each term in the high-fidelity model. Dominant-physics balance-laws can even be determined in different localized spatial and time regions. The clusters can also be identified in a hierarchical way, starting with the most correlated. This strategy leads to discovery of a multi-fidelity hierarchy of balance laws. These models can then be additional reduced using dimension reduction techniques, giving a basis of **empirical orthogonal functions (EOFs)**. These three steps serve as preprocessing for the final step. The final step to describe the framework is **Building Models or Analysis** using the discovered dominant physics clusters and balance laws. After the EOFs are determined, they can serve as a situation aware basis from which to build a surrogate model. EOFs serve as the starting point of the data-driven process for methods such Galerkin ROM and other statistical models. Observable or a hybrid of data can be used to construct models as for validation analysis.

Automated Discovery of DOminaNt physics Informed Surrogates (ADDONIS) Framework for Improving Water Cycling Predictability

Suggested Partners/Experts (Optional)

- Max Gunzburger
- Steven Brunton
- Nathan Kutz
- Jared Callaham
- Peter Caldwell
- Rob Jacobs
- Kate Evans

Bibliography

[1] K. Aadithya, P. Kuberry, B. Paskaleva, P. Bochev, K. Leeson, A. Mar, T. Mei and E. Keiter, "Data-driven Compact Models for Circuit Design and Analysis," in *PMLR*, 2020.

[2] J. Hanson, P. Bochev and B. Paskaleva, "Learning compact physics-aware delayed photocurrent models using dynamic mode decomposition," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* 2020.

[3] Y. Guan, C. Sampson, J. D. Tucker, W. Chang, A. Mondal, M. Haran and D. Sulsky, "Computer model calibration based on image warping metrics: an application for sea ice deformation," *Journal of Agricultural, Biological and Environmental Statistics,* vol. 24, pp. 444-463, 2019.

[4] K. Pieper, K. C. Sockwell and M. Gunzburger, "Exponential time differencing for mimetic multilayer ocean models," *Journal of Computational Physics,* vol. 393, p. 108900, 2019.

[5] K. C. Sockwell, "Mass Conserving Hamiltonian-Structure-Preserving Reduced Order Modeling for the Rotating Shallow Water Equations Discretized by a Mimetic Spatial Scheme," *Florida State University,* 2019.

[6] K. C. Sockwell, K. Peterson, P. Kuberry, P. Bochev and N. Trask, "Interface Flux Recovery coupling method for the ocean--atmosphere system," *Results in Applied Mathematics,* vol. 8, p. 100110, 2020.

[7] K. Peterson, P. Bochev and P. Kuberry, "Explicit synchronous partitioned algorithms for interface problems based on Lagrange multipliers," *Computers & Mathematics with Applications,* vol. 78, pp. 459-482, 2019.

[8] J. L. Callaham, J. V. Koch, B. W. Brunton, J. N. Kutz and S. L. Brunton, "Learning dominant physical processes with data-driven balance models," *arXiv : arxiv.org/abs/2001.10019,* 2020.