DARPA ACTM Milestone 1 Report

Quantification of extreme weather events and their future changes using Physics-Informed DeepONet modeling and functional priors

Agreement No. HR00112290029

PI: T. Sapsis (MIT), G. Karniadakis (Brown U), R. Leung (PNNL)

January 27, 2022

1 Introduction

The governing equations of climate and weather prediction models are chaotic, non-linear and multi-scale in nature. This is in part due to the underlying turbulent fluid motions which are strongly coupled with other multi-physics processes such as convection, particle transport etc. Given the ubiquitous effects of weather and climate dynamics on civilization, the ability to observe, measure, reconstruct and predict the evolution of weather and climate systems is of utmost importance. *Exact* computations themselves are prohibitively expensive due to the large degrees of freedom that are active in the climate system.[1] Therefore, a wide range of critically needed subgrid physics parameterizations have been developed to simulate the physical processes and their interactions with atmospheric circulations.[2] Within computation, subgrid physics parameterizations are used to simulate subgrid-scale processes, while processes at or larger than the model grid-scale are explicitly resolved. As uncertainty in subgrid physics parameterizations is a major source of uncertainty in modeling climate change, recent work has used data-driven methods to parameterize the subgrid physics used in coarse-scale climate models to improve the fidelity of such models and reduce uncertainty in climate projections (see references in[3]).

2 Hybrid models

The advent of machine learning (ML) has opened a new era of hybrid physics and data-driven models for simulating the state of a system.[4–6] Within this approach, part of the solution is provided by solving a physics based equation and the remaining physical processes (sometimes unknown) are emulated by a machine learning technique. In climate modeling, increasing efforts have been invested to develop ML parameterizations (or surrogate models) of subgrid physics, but online use of an ML-based parameterization for climate simulations is not always robust[7].[8] Complementing the development of ML parameterizations, recent efforts to model the discrepancy or residual between low-resolution simulations and global reanalysis (treated as observations) or high-fidelity cloud resolving simulations using ML have shown some promises in correcting for biases in low-resolution atmospheric models[3].[9] As the correction acts as an ML column parameterization, this bias correction approach is analogous to the use of ML parameterizations for hybrid modeling. Previous explorations of the bias correction approach are limited by the brute-force use of AI/ML techniques that have been developed for physics-agnostic tasks with moderate generalization capabilities. This project will improve the bias correction approach by using recently developed, operator based and physics informed ML ideas with improved generalization and uncertainty quantification capabilities compared with physics agnostic methods to improve the fidelity of low-resolution atmospheric models for projecting future changes in storms. The key ingredient is to include memory of historical evolution of the system, allowing non-local representations by which the missing physical processes, referred as closure terms, can be parameterized.

In sum, this project will develop a hybrid model of the atmosphere by using ML techniques to model the bias correction in a low-resolution (also referred as low-fidelity) atmosphere model to reproduce the behaviors of a high-quality global reanalysis and/or a high-fidelity model (global convection-permitting model) that resolve cumulus convection. The hybrid

model is based on the atmospheric component of the Energy Exascale Earth System Model (E3SM)[10] called E3SM Atmosphere Model (EAM)[11] at low resolution, and the high-fidelity model is based on the Simple Cloud-Resolving EAM (SCREAM).[12] The approaches to develop the hybrid model are described in the subsections below. Our approaches will first be tested using a quasi-geostrophic (QG) model, which approximates the atmospheric flow in the mid-to-high latitudes where Coriolis force is strong. We will then employ data from climate models, as described in detail in the next sections.

2.1 Metrics for Hybrid Models

We are going to develop metrics appropriate for evaluating the improvements obtained with the hybrid models:

- One class of metrics will measure to what degree the coarse-scale hybrid model is able to correctly reproduce the spatio-temporal statistics of the large scale features (spectrum in space and time, probability density functions of large scale features) of the observations or high-fidelity simulations used as the reference data for correcting the coarse-scale model, as well as the behavior of these statistics as parameters of the system vary. Specifically, we plan to utilize L_1 and L_2 metrics of the spatial spectrum and temporal autocorrelation functions for specific modes of interest. We will also employ metrics appropriate for measuring performance for extreme events, i.e. tails of the non-Gaussian probability density functions, such as Kullback–Leibler divergence.

- The above metrics will also be used to measure to what degree the coarse-scale model without the ML corrections is able to reproduce the large-scale features of the observations to quantify improvements by the hybrid model.

- The second class of metrics will focus on quantifying the ability to accurately reproduce statistics and their dependence on system parameters for other quantities of interest that depend on the model variables, e.g. heat fluxes for the QG model or atmospheric circulation for the climate datasets.

- Uncertainty quantification metrics that will quantify the statistical characteristics of the generalization error and how this propagates over time through the hybrid model.

2.2 Hybrid Modeling Using Operator Learning

As discussed above, a complimentary approach to surrogate modeling or ML parameterization is to bias correct the approximate solutions from low-fidelity models at regular intervals. In this work, we use data-assimilation in low-fidelity (coarse) simulations to improve their predictions such that they are comparable to more expensive high-fidelity simulations. Using ML, the correction terms from the data assimilation will be predicted using the simulated state variables as predictors. The recently proposed DeepONet[13] is able to satisfy this need for learning the correction terms that map between coarse grid solution and the solution from a high-fidelity (or fine grid) model. DeepONet is a deep neural network approach that maps infinite-dimensional functional spaces rather than finite dimensional vector spaces. The

model contains two deep neural networks. One encodes the input function for multiple initial conditions at given points (coarse grid data) and another encodes the output function.

Our application is based on the following protocol. We have a reference field ($\boldsymbol{u_{ref}}$) available globally over a time period. The governing dynamical equation of the reference field is known

$$\partial_t \boldsymbol{u_{ref}} = \mathcal{N}(\boldsymbol{u_{ref}}) + \boldsymbol{f} \tag{1}$$

where $\mathcal{N}$ is a non-linear operator (e.g. a high-fidelity climate system), $\boldsymbol{f}$ is an external force or driving mechanism, along with a set of boundary conditions. The reference field is derived from a high-fidelity simulation of this system. The initial conditions for the reference flow can also be evolved using a computationally inexpensive simulation over a coarse grid ($\tilde{u}$), at the same locations as the reference solution i.e.

$$\partial_t \tilde{\boldsymbol{u}} = \mathcal{N}(\tilde{\boldsymbol{u}}) + \mathcal{S}(\tilde{\boldsymbol{u}}) + \tilde{\boldsymbol{f}} \tag{2}$$

where $\tilde{\boldsymbol{u}}$ is the solution over a coarse grid. However, certain physical processes may operate over scales smaller than this grid and affect the overall dynamics of the system. These are often accounted for using a closure term or sub-grid scale model $\mathcal{S}$ which parameterizes dynamically important processes that operate at scales smaller than the grid size. The choice of the subgrid model often dictates the level of accuracy and types of physics captured.

The solution of $\tilde{\boldsymbol{u}}$ is biased due to the lack of the resolution compared to $\boldsymbol{u_{ref}}$. The bias can be corrected via a Newton relaxation approach (also referred to as "Nudging") by adding a correction term on the righthand side of Eq(2):

$$\partial_t \tilde{\boldsymbol{u}} = \mathcal{N}(\tilde{\boldsymbol{u}}) + \mathcal{S}(\tilde{\boldsymbol{u}}) + \tilde{\boldsymbol{f}} + \dot{\boldsymbol{R}} \tag{3}$$

where $\dot{\boldsymbol{R}}$ denotes the nudging tendency which can be treated as an error correction term at each model step and grid point. If $\boldsymbol{u_{ref}}$ is known, $\dot{\boldsymbol{R}}$ can be estimated by:

$$\dot{\boldsymbol{R}} = -\frac{\tilde{\boldsymbol{u}} - \boldsymbol{H}\boldsymbol{u_{ref}}}{\tau} \tag{4}$$

where τ is defined as a relaxation time scale that controls the strength of the correction with $\dot{\boldsymbol{R}}$. $\boldsymbol{H}$ is an operator that maps $\boldsymbol{u_{ref}}$ to the space of $\boldsymbol{u}$. We note that $\boldsymbol{u_{ref}}$ can represent any solution determined to be of higher fidelity than the default solution to Eq(2). Common examples for $\boldsymbol{u_{ref}}$ are observations, reanalyses products, or model simulations run at higher resolution or with more advanced physical parameterizations. In addition, $\dot{\boldsymbol{R}}$ accounts for the any physical processes included in $\boldsymbol{u_{ref}}$ that are absent in the default configuration. The nudging approach with Eq(3) and Eq(4) has been shown to faithfully reproduce complex physics in a range of systems using fewer degrees of freedom than available in high-fidelity simulations.[14–16]

In this project, we aim to use DeepONet to provide the correction term $\dot{\boldsymbol{R}}$ at each time-step using only $\tilde{\boldsymbol{u}}$, instead of requiring a high-fidelity simulation for each low-resolution counterpart. i.e.

$$\dot{\boldsymbol{R}} = -Q(\tilde{\boldsymbol{u}}) \tag{5}$$

where Q is a filtering operator that reconstructs the $\dot{\boldsymbol{R}}$ with the information of $\tilde{\boldsymbol{u}}$. Since the goal is to use partial data to reconstruct the full system configuration, we will seek to answer two questions in a systematic approach. The first is the quantity of information needed to achieve a given degree of reconstruction. The second relates to the quality or type of data needed for the reconstruction.

2.3 Hybrid modeling using Physics Informed Neural Operators

In addition to the available data from the reference solution we plan to employ appropriate physical constraints, such as conservation of mass, humidity, energy, etc. These will employed to improve the training error for the operator Q. To merge the information from the physical constraints with the data we plan to utilize the Physics-Informed Deep-ONet. The main idea of this method is to significantly enhance training and accuracy by including some of the governing equations and other physical constraints in the loss function during the training stage, that follows the PINNs paradigm, where physics and data both contribute to training the network. This will help us to analyze the functional spaces of atmospheric states, i.e., temperature, moisture, horizontal winds, surface pressure, etc.. It can seamlessly integrate multifidelity/multimodality experimental data with mathematical physics. We will begin by investigating the two-layer quasi-geostrophic (QG)[17,18] as a beta model. Upon analyzing the simulation results, we will modify/enhance our methodology accordingly and then move towards the E3SM model. The E3SM (Energy Exascale Earth System Model)[19] is the Earth system model developed by the Department of Energy for investigating energy-relevant science and climate prediction. The E3SM focuses on compute-intensive frontiers in climate science; one objective is to develop a new GCPM (Global Convection-Permitting Models) called the Simple Cloud-Resolving E3SM Atmosphere Model (SCREAM)[20] . SCREAM consists solely of non-hydrostatic fluid dynamics, a turbulence/cloud fraction scheme, a microphysics scheme, a radiation scheme, an energy fixer, and prescribed-aerosol functionality. This information can be very well be coded into the physics informed DeepONet, and hence we aim to aid the E3SM model by creating the auto-correcting nudging tendencies. This will help the model to predict extreme weathers and tipping points even when its run on a coarse-grid simulation.

2.3.1 Physical Constraints for energy transfers

Here we give an example of such physical constraint involving energy transfer due to nonlinear terms in fluid flows. A data-informed representation of the large-scale quantities of a turbulent climate model must accurately capture the energy transfers between the resolved and unresolved energetic modes of the system. The importance of these interactions are attested in the strong forward and backward energy cascades geostrophic flows exhibit. Furthermore, any turbulent system by definition experiences persistent instabilities, which require some of the energetic modes of the system to be unstable. Hence, the only way for such a system to equilibrate, is for unstable modes to transfer energy to the stable ones. As a result, the numerical stability of a data-driven model is strongly tied to the ability of the model to accurately depict such interactions. One way to reinforce the system to respect these mechanisms is to add physical constraints during training. One such constraint arises

from the assumption that the advection term of turbulent flows is energy-preserving, i.e. it cannot change the total kinetic energy of the system. For a general incompressible flow field $\boldsymbol{u}$, one can write the integral condition

$$\int_{\Omega} \boldsymbol{u} \cdot (\boldsymbol{u} \cdot \boldsymbol{\nabla}) \boldsymbol{u} \mathrm{d}\boldsymbol{x} = \int_{\partial\Omega} \frac{1}{2} (\boldsymbol{u} \cdot \boldsymbol{u}) \boldsymbol{u} \cdot \boldsymbol{n} \mathrm{d}\boldsymbol{x} \tag{6}$$

where Ω is the spatial flow regime and $\boldsymbol{n}$ is an outwards-pointing unitary-vector normal to $\partial\Omega$. By introducing the separation of $\boldsymbol{u}$ into its large-scale mean and small-scale perturbations $\boldsymbol{u}'$ this conditions can be written as a constraint for the predicted $\boldsymbol{u}$ during training (when small-scale data will be available). Additional constraints like conservation of mass or momentum can be used in the same context when appropriate.

2.3.2 A simple example

Let us consider a pedagogical example - the Burgers equation:

$$u_t(x,t) + uu_x(x,t) - \nu u_{xx}(x,t) = 0, \qquad x \in [-1,1], t \in [0,1], \tag{7}$$

$$u(0,t) = u(1,t) = 0, \qquad t \geq 0, \tag{8}$$

$$u(x,0) = -\sin(\pi x), \quad x \in [-1,1], \tag{9}$$

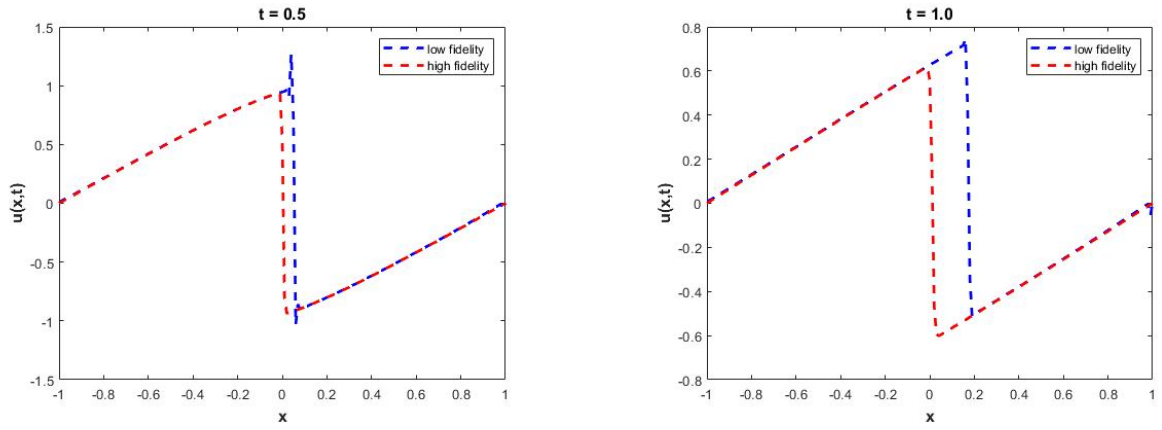
where $\nu = (0.01)/\pi$.

The above equation is solved using a finite difference method in coarse-grid resolution as well as fine-grid resolution. The coarse-grid simulation is a low-fidelity solution and has a lot of error compared to the one solved in a finer-grid resolution, which is a high fidelity solution (which is closer to the exact solution). Figures 1 and 2 represent the scenarios for two different finite difference[21] coarse-grid (40-points and 200-points) simulation compared to the finite difference fine-grid simulation (2000 points) projected on the coarse-grid (40-points and 200-points). For the coarse-grid low-fidelity simulation, 40 and 200 points were used in the space domain and for the fine-grid simulation 2000 points were used in the space domain. For both the cases 1000 points were used in the time domain. It can be clearly seen that, coarser the grid the more information about the solution is lost.

This lost information could be very vital to the prediction of the behaviour of the model. So our main objective is to run the coarse resolution model and use the information from the fine resolution model and learn the correction factor so that the coarse model can be used to give the similar level of accuracy like the fine resolution. Mathematically, it can be expressed as

$$\delta_t \tilde{\mathbf{u}} + 0.5(\tilde{\mathbf{u}}_{\mathbf{x}})^2 - \nu \tilde{\mathbf{u}}_{\mathbf{xx}} + \boldsymbol{\lambda} ||\tilde{\mathbf{u}} - \mathbf{u}_{\mathbf{ref}}^{(\mathbf{p})}||^2 = 0, \tag{10}$$

where $\tilde{\mathbf{u}}$ is the coarse resolution output, $\mathbf{u}_{\mathbf{ref}}^{(\mathbf{p})}$ is the projection of the fine resolution output in the coarse grid and $\boldsymbol{\lambda}$ is the correction factor. $\boldsymbol{\lambda} = (\lambda_0, \lambda_b, \lambda_r)$, where λ_0 is the correction term at the initial points, λ_b is the correction term at the boundary points, and λ_r is the correction term at the interior points. When training, $\boldsymbol{\lambda}$ can be introduced as a fixed term (inefficient) or it can be introduced as a self adaptive variable corresponding to the training/residual points (more-effective). Making the correction factor self-adaptive is with the aim of minimize the term $||\tilde{\mathbf{u}} - \mathbf{u}_{\mathbf{ref}}||^2$ while maximizing $\boldsymbol{\lambda}$. This ensures proper learning of the physics involved.



(a) (b)

Figure 1: Comparison between finite-difference method solutions at time (a) t=0.5 and (b) t=1.0, solved in coarse-grid (200-points) and fine-grid (2000-points) projected on coarse-grid (200-points)

It can be also seen from the Figures 1 and 2, that the the correction factor will correspond to function as it will be varying based on the location of the point in space and time. Hence, our plan is to employ DeepONnet and represent the nudging tendencies as an operator.

2.4 Loss functions and other considerations for training

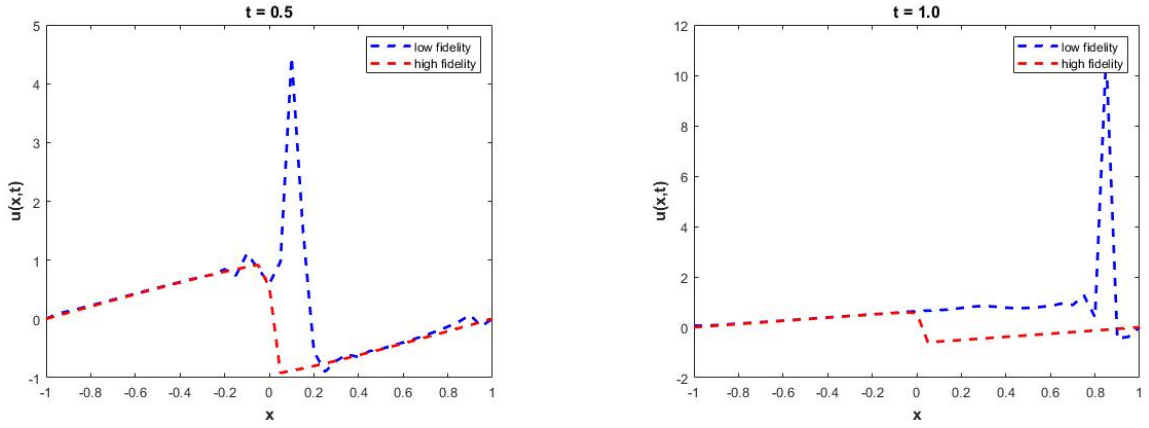
Since we will be using neural networks for approximations, the convergence depends a lot on how we define our loss function. In most applications, the L_2 norm for defining the loss term is employed. The L_2 norm for a function ϕ is given by,

$$||\phi||_{L_2}^2 \equiv \phi \cdot \phi \equiv \langle \phi \mid \phi \rangle \equiv \int |\phi(x)|^2 dx. \tag{11}$$

In this study, we also want to investigate different norms or a combination of norms in relation to the speed of convergence. It has been shown theoretically the loss functions defined in terms of Sobolev norm guarantees the convergence of neural network to a solution of PDEs in the corresponding Sobolev spaces.[22] For example, the H_1 norm for a function ϕ is given by

$$\|\phi\|_{H_1} = \sqrt{\|\phi\|_{L_2}^2 + \|\phi'\|_{L_2}^2}.$$

So it might be more effective to use the H_1 norm for a faster convergence. Also if the initial condition exhibits high oscillations (e.g., assuming the input space is a Gaussian random field with small correlation length) it might be better to utilize the H_1 norm. As we further investigate the problem, we will run and analyze different mixed norms to evaluate how it effects the convergence of the correction factor.



(a) t=0.5 (b) t=1.0

Figure 2: Comparison between finite-difference method solutions at time (a) t=0.5 and (b) t=1.0, solved in coarse-grid (40-points) and fine-grid (2000-points) projected on coarse-grid (40-points)

2.5 Uncertainty quantification using functional priors

After this, we plan to analyze the uncertainty quantification and quality of training data. We plan to use the physics-informed Generative Adversarial Networks (PI-GAN)[23] to learn a functional prior from data (nudging tendencies) and physics. We will employ the Hamiltonian Monte Carlo (HMC) method to estimate the posterior in the latent space of PI-GANs. Here, we will use the output of the PI-GAN as an input to DeepONet. We will then assess the quality of training data. In particular we aim to develop a general formulation, inspired by recent developments on active learning algorithms.[24]

Learning functional priors and functional posteriors extends the Bayesian framework to discover very informative priors using existing data, (partial) physics, and GANs for semi-supervised learning. Unlike Bayesian neural networks, which taxes computational resources heavily due to the high-dimensionality (every weight and bias is a random parameter that needs a prior), in Functional priors and posteriors we work in the latent space of GANs, so the dimensionality of the stochasticity is manageable (around 50). Specifically, for the "beta" QG model, we will generate training data by varying the model parameters in over a certain range and run the model to obtain sets of simulated data that can be use in obtaining informative functional priors. Having gained experience with the beta model, we will then apply to the coarse simulation data of the climate model. In this case, we will consider different complete or gappy training data for some or all of the state variables. The main premise here that in addition to quantifying uncertainty, we will also be able to make predictions using a minimalistic set of new data to obtain the functional posterior, given that we have a head start with the functional prior.

2.6 Summary for hybrid modeling algorithms

We have broadly summarized our approach into the following sub-tasks involving hybrid modeling:

(i) Use the QG Model (beta phase) and E3SM Model (first-phase climate data set) for simulating data for coarse grid simulation and learn the nudging tendency by comparing with fine-grid simulation by by utilizing Physics Informed DeepONet.

(ii) Employ DeepONet to predict the correction term as an operator.

(iii) Employ physical principles using Physics-Informed DeepONet and adaptive weights.

(iv) Analyze the properties of different norms with respect to the loss function.

(v) Perform uncertainty quantification of the proposed method and long term climate prediction.

3 Datasets

3.1 Dataset for beta version: tipping points for extreme events in a QG model

The first stage of this study focuses on data-driven uncertainty quantification of different physical regimes, generated by a simple climate model. In more detail, reduced-order models, able to capture the parametric transition between statistical regimes of the reference climate models, are investigated. Of particular interest are the tails in probability density functions of large-scale flow features and their transition from sub-Gaussian behaviour to the appearance of heavy-tails. To that end, a two-layer quasi-geostrophic (QG)[17] model is employed for the velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$. The turbulent flow is described by the evolution equation

$$\frac{\partial q_j}{\partial t} + \boldsymbol{v}_j \cdot \boldsymbol{\nabla} q_j + \left(\beta + k_d^2 U_j\right) \frac{\partial \psi_j}{\partial x} = -\delta_{2,j} r \Delta \psi_j - \nu \Delta^s q_j + Q\left(q_j - q_j^{ref}\right), \tag{12}$$

where $j = 1, 2$ corresponds to the upper and lower layer respectively, q_j. A mean zonal flow of intensity $U_1 = U$ and $U_2 = -U$ is imposed on each layer respectively. The two layers are assumed to have the same width, k_d denotes the deformation radius, r the bottom-drag coefficient and β is the beta-plane approximation parameter. The term Q is a nudging term that corrects the coarse-scale solution based on the fine-scale reference solution. When running fine-scale simulations this term is equal to zero. The potential vorticity (PV) q_j and corresponding streamfunction ψ_j are related via the inversion formulae

$$q_j = \Delta \psi_j + \frac{k_d^2}{2}\left(\psi_{3-j} - \psi_j\right), \quad j = 1, 2. \tag{13}$$

The velocity field can be written as $\boldsymbol{v}_j = (U_j, 0) + \boldsymbol{v}'_j$, where it is decomposed into a zonal mean and a fluctuating shear flow, with $\boldsymbol{v}'_j = (-\partial \psi_j / \partial x, \partial \psi_j / \partial y)$. For a more physical interpretation of the flow, the barotropic PV $q_t = (q_1 + q_2)/2$ and the the baroclinic PV

$q_c = (q_1 - q_2)$ are defined respectively. A typical snapshot of flows generated by this model can be seen in figure 3.

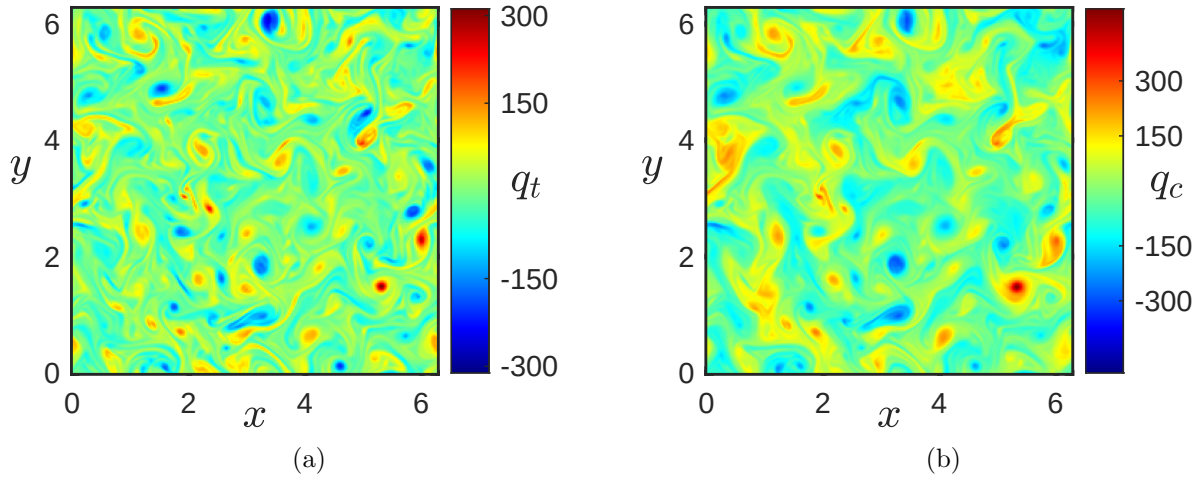


Figure 3: Typical snapshots (vorticity fields) of the barotropic and baroclinic mode for baroclinic ocean turbulence at high latitudes.

Changing the parameters of this model allows for the transition of large-scale statistics to different probabilistic regimes.[25] The parameter values for two such cases are shown in Table 1. An example of this transition is depicted in figure 4, where the marginal pdfs for the leading modes $\hat{\psi}_{(1,0)}, \hat{\psi}_{(0,1)}$ of the barotropic streamfunction $\psi = (\psi_1 + \psi_2)/2$ are presented and contrasted with a best-fitted Gaussian.

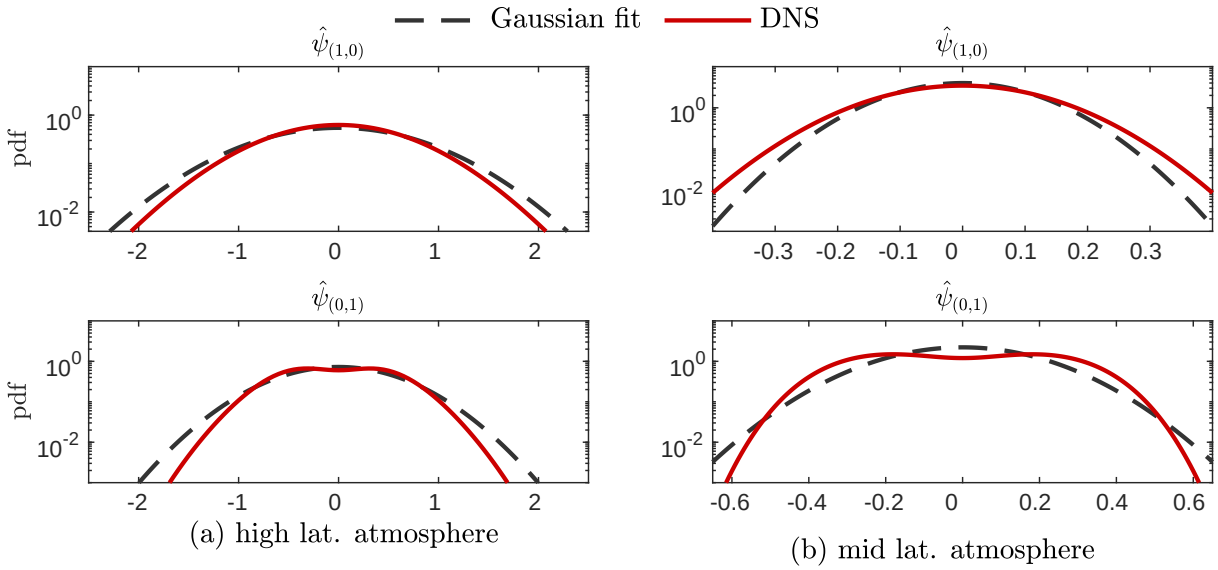


Figure 4: Marginal pdfs of the two leading modes of the barotropic streamfunction, produced via direct numerical simulations (red) for two different test-cases, and their corresponding Gaussian fit (black).

regime	β	k_d	U	r	ν	s
atmosphere, high lat.	1	4	0.2	0.2	1×10^{-13}	4
atmosphere, mid lat.	2	4	0.2	0.1	1×10^{-13}	4

Table 1: Parameter values for different atmosphere regimes.

Despite the relatively simple structure of the model, the existence of intermittent energy cascades magnify the importance of small-scale feedbacks to large-scale flow features. This phenomenon adds a major difficulty in accurately predicting the statistics of the flow due to inadequate resolution of small-scale interactions. Furthermore, conventional Gaussian-closure schemes lack the ability to correctly capture the emerging non-Gaussian statistics. Hence, reduced-order models that appropriately model the energy transport mechanism between modes of the system are essential.

3.2 Climate Datasets

3.2.1 Climate datasets for Phase 1

We will employ version 2 of the Energy Exascale Earth System Model (E3SMv2) to create climate data. The lower-resolution configuration with approximately 1° (∼110 km) resolution in horizontal and 72 layers in vertical is the target model (also referred as coarse resolution model) to be improved with the machine learning approach, while a high resolution (order 3 km grid spacing) configuration will be used to generate the high-fidelity reference that the coarse resolution model will be used to produce corrections for machine learning to improve its skill. The high-resolution E3SM experiment is computationally expensive as increasing horizontal resolution from 110 km to 3 km results in 37x more columns in both x and y directions, which requires approximately 6400x more computational resources for a single simulation. Therefore, while that is underway, we will start with a nudging setup that uses a high quality global reanalysis as the high fidelity reference for producing the nudging input. Reanalysis products are currently the best estimation of past weather and climate achieved by a high fidelity model in the operational weather forecast centers. The ERA5 reanalysis[26] from the European Centre for Medium-Range Weather Forecasts (ECMWF) are used for generating nudging data. The 3-hourly ERA5 reanalysis data on a 0.25^o grid are spatially remapped to the cubed-sphere grid and 72 model layers used by E3SMv2, following the method used in the Community Earth System Model Version 2 (CESM2; `https://ncar.github.io/CAM/doc/build/html/users_guide/physics-modifications-via-the-namelist.html#nudging`). Topographical differences between E3SMv2 and the reanalysis model are taken into account during the vertical interpolation. The validation of the ERA5 reanalysis for nudging has been documented for the predecessor version of E3SM,[16] which shows that the high-quality ERA5 reanalysis improves the performance of nudging in E3SM. We also note that the implementation of nudging is already available in the E3SMv2 code base, and ERA5 reanalysis for nudging input datasets has been appropriately constructed for this project.

As outlined in the previous work, there are a number of choices that need to be made with respect to the nudging,[14,15,27–30] as well as to which state variables and fluxes are desired as

input to learn the nudging tendencies.[31,32] Table 2 lists groups of simulations that will be conducted to provide states and nudging tendencies [i.e. $\dot{\boldsymbol{R}}$ in Eq(4) and Eq(5)] used for the machine learning procedures. All the simulations involved active atmosphere and land but used prescribed sea surface temperature (SST) and sea ice extension, following the protocol from the Atmospheric Model Intercomparison Project.[33] The SST and sea ice extension used in this study are weekly data from the National Oceanic and Atmospheric Administration (NOAA) Optimum Interpolation (OI) analysis.[34] Other external forcings, including volcanic aerosols, solar variability, concentrations of greenhouse gases, and anthropogenic emissions of aerosols and their precursors, were prescribed following the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project-Phase 6 [CMIP6].[35–37] Emissions of aerosols and their precursor gases of the year 2010 are used to represent the present-day (PD) condition.

Table 2: List of nudged EAMv2 simulations. Nudging was applied at each physics time step (30 minutes)

Group number	Simulation short name	Nudged variables	Constraining data and frequency	Nudging relaxation time scale(τ)	Dataset ($\tilde{\boldsymbol{u}}$ and $\dot{\boldsymbol{R}}$)
1	CLIM	None	N/A	N/A	U, V, T, Q
2	NDG_T	T	ERA5 (3 hr)	12 hr	R_T
2	NDG_Q	Q	ERA5 (3 hr)	12 hr	R_Q
2	NDG_UV	U, V	ERA5 (3 hr)	12 hr	R_U, R_V
2	NDG_UVT	U, V, T	ERA5 (3 hr)	12 hr	R_U, R_V, R_T
2	NDG_UVQ	U, V, Q	ERA5 (3 hr)	12 hr	R_U, R_V, R_Q
2	NDG_UVTQ	U, V, T, Q	ERA5 (3 hr)	12 hr	R_U, R_V, R_T, R_Q
3	NDG_T	T	ERA5 (3 hr)	6 hr	R_T
3	NDG_Q	Q	ERA5 (3 hr)	6 hr	R_Q
3	NDG_UV	U, V.	ERA5 (3 hr)	6 hr	R_U, R_V
3	NDG_UVT	U, V, T	ERA5 (3 hr)	6 hr	R_U, R_V, R_T
3	NDG_UVQ	U, V, Q	ERA5 (3 hr)	6 hr	R_U, R_V, R_Q
3	NDG_UVTQ	U, V, T, Q	ERA5 (3 hr)	6 hr	R_U, R_V, R_T, R_Q
4	NDG_T	T	ERA5 (3 hr)	3 hr	R_T
4	NDG_Q	Q	ERA5 (3 hr)	3 hr	R_Q
4	NDG_UV	U, V	ERA5 (3 hr)	3 hr	R_U, R_V
4	NDG_UVT	U, V, T	ERA5 (3 hr)	3 hr	R_U, R_V, R_T
4	NDG_UVQ	U, V, Q	ERA5 (3 hr)	3 hr	R_U, R_V, R_Q
4	NDG_UVTQ	U, V, T, Q	ERA5 (3 hr)	3 hr	R_U, R_V, R_T, R_Q

The first group of simulation is a free-running baseline simulation referred to as CLIM, from which the prognostic variables including wind (U, V), temperature (T), and specific humidity (Q) are archived to provide the information of large-scale dynamics and thermodynamics from the low-fidelity (coarse-resolution) E3SM model. The second to fourth groups of simulations will be conducted by nudging the E3SM toward ERA5 reanalysis with three different setups for nudging timescale (τ). There are five simulations in each group with different combinations of nudged prognostic variables. The difference in simulation setup is whether nudging applied

- only to the temperature (labeled "_T"),
- only to the humidity (labeled "_Q"),

- only to the horizontal winds (labeled "_UV"),
- to both winds and temperature (labeled "_UVT"),
- to both winds and humidity (labeled "_UVQ"), and
- to winds, temperature, and humidity (labeled "_UVTQ")

In a nudged simulation, a nudging tendency term in the form of Eq. (4) is calculated for each of the prognostic variables. The pink boxes in Figure 5 illustrate where the nudging-related calculations occur in the atmospheric component of E3SM. The prognostic variables with the values saved before the radiation calculation were compared with the ERA5 reanalysis to derive the nudging tendency terms. These terms are used to update the model state variables after the physics parameterization suite (before the next call of the dynamical core).

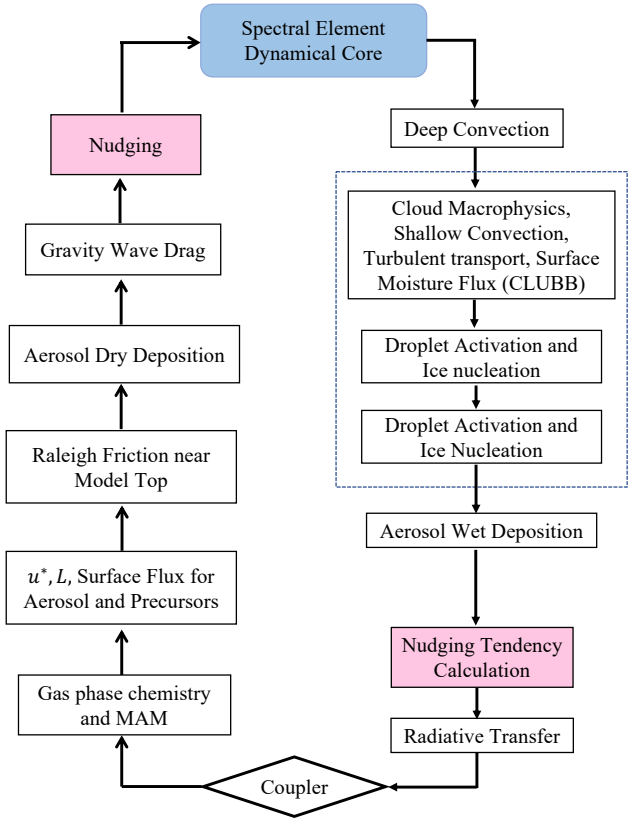


Figure 5: Flowcharts showing the sequence of dynamics and physics calculations within one time step in an E3SMv2 simulation. Pink boxes indicate where the nudging-related calculations occur.

By varying the nudging timescale and the nudged prognostic variables, we can generate a larger swath (ensembles) of states and nudging tendencies (c.f. fifth column in Table 2) that (1) broadens the training data available to the machine learning procedures, and (2) includes the uncertainties of the state and nudging tendencies arising from the nudging strategy in the machine learning for this project. In addition, the surface heat, moisture, and radiative

fluxes in each simulation will also be archived to facilitate the design of machine learning in this project with a better treatment on the large land-surface drift in the nudged training run in previous studies.[31,32]

3.2.2 Climate datasets for Phase 2

Following the same methodology described for generating climate datasets for Phase 1, we will develop climate datasets for Phase 2 by replacing ERA5 with high resolution simulations by the high-fidelity E3SM applied at order 3 km grid spacing. The high-resolution E3SM simulations will be spatially interpolated to the coarse-resolution mesh used by the target model for the nudging simulations to generate correction terms for ML training.

4 Active learning for evaluating and improving data quality

We aim to leverage hybrid models (i.e. machine-learning models) for implementing novel data acquisition techniques, assess the quality of fixed datasets, and determine target simulations for assessing climate tipping points. The hybrid models we aim to explore consist of functional priors obtained using DeepONet, describe before. While Gaussian Process is seen as the "gold standard" in Bayesian experimental design (BED) and optimization (BO), it suffers catastrophically when either data or the dimension of problems becomes large. DeepONet and functional priors do not suffer from such constraints.

4.1 Active Learning

To implement active learning on our various datasets and dynamical systems we have several steps that must be achieved. First we must define suitable functions that map inputs to outputs based on the available data ($\mathbf{y} = f(\mathbf{x})$). This will be followed by implementation of several acquisition functions with DeepONet.

4.2 Identification of appropriate inputs and extreme outputs

1. Define "danger map" definition. Items include setting the time horizon and definition of appropriate observables (e.g. max value at $t = \tau$).
2. Set appropriate definitions of priors for our datasets and dynamical systems (e.g. uniform or Gaussian distributions).
3. Identify appropriate parameter ranges for dynamical systems of interest.

4.3 Uncertainty quantification and acquisition functions

1. Determine the ensemble size for sufficiently expressive uncertainty quantification, $\sigma^2(\mathbf{x})$, and for an accurate mean model $\mu(\mathbf{x})$. An ensemble of trained neural networks is used to determine both of these quantities. This is necessary as there are no guarantees that

a specific choice in ensemble size will be sufficient for expressing any general problem at hand.
2. Determine the impact of the several hyperparameters: training time (epochs), neurons (too little or too many), layers, architectures, activation functions, learning rates and scheduling, among others
3. Apply several previously developed acquisition functions (such as uncertainty sampling, latin hypercube sampling, and the likelihood-weighted sampling approaches) using NNs and assess the performance of these active sampling procedures.
4. Determine the utility of a "batching" technique, where multiple points are chosen for acquisition at each step. Batching will be used to reduce experiment iterations when experiments may be run in parallel.

4.4 Higher-order Acquisition functions

In addition to the previously used acquisition functions, we aim to employ an extension of the Likelihood-Weighted Integrated Variance Reduction (IVR-LW) acquisition function as described in the following:

1. Implement the high-order extension of the Likelihood-Weighted Integrated Variance Reduction (IVR-LW) to GPs. This will determine the optimal samples to best represent the extreme events and *pdf* discontinuities (related to tipping points).

$$\alpha(\mathbf{x}) = k_1 c_0^2 \int_{S_\mathbf{x}} w_1(\overline{\mathbf{x}})\sigma^2(\overline{\mathbf{x}})\mathrm{d}\overline{\mathbf{x}} + \left[\int_{S_\mathbf{x}} \left(k_2 w_2(\overline{\mathbf{x}}) + k_3 \left|\overline{w}_3(\overline{\mathbf{x}})\right|\right)\sigma^2(\overline{\mathbf{x}})\mathrm{d}\overline{\mathbf{x}}\right]^2, \tag{14}$$

where

$$w_1(\mathbf{x}) = \frac{p_\mathbf{x}(\mathbf{x})}{p_\mathbf{y}(f(\mathbf{x}))}, w_2(\mathbf{x}) = \frac{p_\mathbf{x}^2(\mathbf{x}) p_\mathbf{y}'^2(f(\mathbf{x}))}{2p_\mathbf{y}^4(\mathbf{y}(\mathbf{x}))}, \overline{w}_3 = \frac{\nabla^T f(\mathbf{x}) \mathbf{H}_{f(\mathbf{x})} \nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|^4 + k_3\varepsilon} p_\theta(\overline{\theta}). \tag{15}$$

2. Investigate how well the higher-order extension of the IVR-LW acquisition function takes into account the geometry of the the mean model, $\mathbf{y} = f(\mathbf{x})$, as well as the probabilistic properties of the model.

4.5 Quality of Fixed Datasets

We are also interested in determining how one properly uses fixed datasets for training machine learning models, which is extremely relevant for this problem. This is a long standing question in the machine learning literature, yet limited approaches exist. Typically, all available data is used in training, however, using all data is not always appropriate. Static data often contains misleading data points that lead to inaccurate models. Methods for mitigating or eliminating misleading data are thus crucial to appropriately leverage historical data. Here we consider two techniques, an active learning metric for fixed data and functional priors, for mitigating the effects of misleading data and enabling widespread use of historical datasets.

4.5.1 Active learning for fixed training data

Training models on *all* available data does not always lead to the best results. In fact, increasing data does not monotonically lead to better model performance and commonly leads to poor generalization of testing data. This behavior is known as "double-descent",[38] where after finding initially good performance, the model retreats in quality as data increases until eventually returning to good performance as data approaches infinity. Although this may be bypassed with plentiful data, this is not permissible for large-scale systems with limited data. Further, it is not even clear when and where double-descent is occurring for problems where truth data does not exist. We propose several approaches to double-descent mitigation by eliminating misleading data through probabilistic measures popularized in active learning. We present these ideas below and outline various questions to be answered throughout the project.

1. Identify datasets/systems with and without the double descent phenomenon.
2. Create formal development of fixed dataset metrics. Currently, we propose a likelihood-weighted dropout (LW-DO), a likelihood-weighted addin (US-LW-AI), and a combination of both methods (LW-DO + US-LW-AI).
3. Implement these metrics on datasets/systems ranging from low to high dimensions and small to large with NNs and GPs. Assess performance against truth data.
4. Assess the viability of an iterative version of LW-DO + US-LW-AI.
5. Determine stopping criteria or cut-offs for misleading data when truth data is unavailable.

5 Project Summary and Milestones

A summary of the steps and objectives for this project are listed below (Milestones are according to the DARPA excel spreadsheet):

M2: Develop framework for the hybrid models identifying the known and unknown parts along with mathematical approaches.

1. Identify important parameter regimes for the systems considered and characterize the corresponding statistical regime.

2. Develop hybrid framework and related mathematical approaches.

M3: Deliver prepared datasets for use in Phase 1. Report on metrics to be used to compare the benefits of hybrid models over conventional models.

1. Generate a large database of reference data, with time evolution, across a vast range of parameter regimes.

2. Generate a database of coarse grid simulations for the same set of conditions and parameter regimes.

3. Formulate statistical metrics to quantify improvement (such spectrum, probability densities of quantities of interest).

M4: Report on 'beta version' of hybrid model analysis including new mathematical insights, along with insights in data analysis.

1. Identify suitable nonlinear mappings that transform input to outputs based on the QG ('beta') data generated above with DeepONet.

2. Utilize active learning protocols to evaluate quality of data and improve training errors.

M5: Identify potential datasets (and providers) for Phase 2 to address predictability of climate effects at 1-to-3 decade time scales and regional or global spatial extents.

1. See section 3.2.2.

M6: Deliver progress report and preliminary software for the hybrid models/methods.

1. Complete developments of DeepONet and active learning algorithms, i.e. items 1,2 of M4.

2. Improve accuracy using physical constraints and Physics-Informed DeepONet.

3. Determine the effects of various hyper-parameters such as training time, cadence time in reference data, learning rates.

M7: Deliver detailed research plan for Phase 2 to include specifics of data (provider, availability) and expected updates to the model.
M8: Deliver final report on Phase 1 to include data analysis and characterization of benefits of new hybrid models over conventional models. Examples could include ability to characterize tipping points in ways not possible by conventional models, orders of magnitude speedup using AI proxy methods without loss of accuracy, etc.

1. Complete development of Physics-Informed DeepONet, apply to climate datasets and evaluate performance for different physical constraints.

2. Evaluate accuracy in terms of statistical properties for large scale features of interest.

References

[1] Bjorn Stevens, Masaki Satoh, Ludovic Auger, Joachim Biercamp, Christopher S Bretherton, Xi Chen, Peter Düben, Falko Judt, Marat Khairoutdinov, Daniel Klocke, et al. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1):1–17, 2019.

[2] Masaki Satoh, Bjorn Stevens, Falko Judt, Marat Khairoutdinov, Shian-Jiann Lin, William M Putman, and Peter Düben. Global cloud-resolving models. *Current Climate Change Reports*, 5(3):172–184, 2019.

[3] Christopher S Bretherton, Brian Henn, Anna Kwa, Noah D Brenowitz, Oliver Watt-Meyer, Jeremy McGibbon, W Andre Perkins, Spencer K Clark, and Lucas Harris. Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. 2021.

[4] Zhong Yi Wan, Boyko Dodov, Christian Lessig, Henk Dijkstra, and Themistoklis P Sapsis. A data-driven framework for the stochastic reconstruction of small-scale features with application to climate data sets. *Journal of Computational Physics*, page 110484, 2021.

[5] Alexis-Tzianni G Charalampopoulos and Themistoklis P Sapsis. Machine-learning energy-preserving nonlocal closures for turbulent fluid flows and inertial tracers. *arXiv preprint arXiv:2102.07639*, 2021.

[6] D Eeltink, H Branger, C Luneau, J Kasparian, TS van den Bremer, and TP Sapsis. Nonlinear wave evolution with data-driven breaking. 2021.

[7] Noah D Brenowitz, Tom Beucler, Michael Pritchard, and Christopher S Bretherton. Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, 77(12):4357–4375, 2020.

[8] Noah D Brenowitz, Brian Henn, Jeremy McGibbon, Spencer K Clark, Anna Kwa, W Andre Perkins, Oliver Watt-Meyer, and Christopher S Bretherton. Machine learning climate model dynamics: Offline versus online performance. *arXiv preprint arXiv:2011.03081*, 2020.

[9] Oliver Watt-Meyer, Noah D Brenowitz, Spencer K Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W Andre Perkins, and Christopher S Bretherton. Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, 48(15):e2021GL092555, 2021.

[10] Jean-Christophe Golaz, Peter M Caldwell, Luke P Van Roekel, Mark R Petersen, Qi Tang, Jonathan D Wolfe, Guta Abeshu, Valentine Anantharaj, Xylar S Asay-Davis, David C Bader, et al. The doe e3sm coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, 11(7):2089–2129, 2019.

[11] PJ Rasch, S Xie, P-L Ma, W Lin, H Wang, Q Tang, SM Burrows, P Caldwell, K Zhang, RC Easter, et al. An overview of the atmospheric component of the energy exascale earth system model. *Journal of Advances in Modeling Earth Systems*, 11(8):2377–2411, 2019.

[12] Peter Martin Caldwell, Christopher Ryutaro Terai, B Hillman, Noel D Keen, P Bogenschutz, Wuyin Lin, Hassan Beydoun, M Taylor, Luca Bertagna, AM Bradley, et al. Convection-permitting simulations with the e3sm global atmosphere model. *Journal of Advances in Modeling Earth Systems*, 13(11):e2021MS002544, 2021.

[13] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[14] K. Zhang, H. Wan, X. Liu, S. J. Ghan, G. J. Kooperman, P.-L. Ma, P. J. Rasch, D. Neubauer, and U. Lohmann. Technical Note: On the use of nudging for aerosol-climate model intercomparison studies. *Atmospheric Chemistry and Physics*, 14(16):8631–8645, 2014.

[15] Jian Sun, Kai Zhang, Hui Wan, Po-Lun Ma, Qi Tang, and Shixuan Zhang. Impact of nudging strategy on the climate representativeness and hindcast skill of constrained eamv1 simulations. *Journal of Advances in Modeling Earth Systems*, 11(12):3911–3933, 2019.

[16] Shixuan Zhang, Kai Zhang, Hui Wan, and Jian Sun. Further improvement and evaluation of nudging in the e3sm atmosphere model version 1 (eamv1). *Geosci. Model Dev. Discuss.*, 2022. Preprint under review for GMDD.

[17] Rick Salmon. *Lectures on geophysical fluid dynamics*. Oxford University Press, 1998.

[18] Geoffrey K. Vallis. *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation*. Cambridge University Press, 2006.

[19] Jean-Christophe Golaz, Peter M. Caldwell, Luke P. Van Roekel, Mark R. Petersen, Qi Tang, Jonathan D. Wolfe, Guta Abeshu, Valentine Anantharaj, Xylar S. Asay-Davis, David C. Bader, Sterling A. Baldwin, Gautam Bisht, Peter A. Bogenschutz, Marcia Branstetter, Michael A. Brunke, Steven R. Brus, Susannah M. Burrows, Philip J. Cameron-Smith, Aaron S. Donahue, Michael Deakin, Richard C. Easter, Katherine J. Evans, Yan Feng, Mark Flanner, James G. Foucar, Jeremy G. Fyke, Brian M. Griffin, Cécile Hannay, Bryce E. Harrop, Mattthew J. Hoffman, Elizabeth C. Hunke, Robert L. Jacob, Douglas W. Jacobsen, Nicole Jeffery, Philip W. Jones, Noel D. Keen, Stephen A. Klein, Vincent E. Larson, L. Ruby Leung, Hong-Yi Li, Wuyin Lin, William H. Lipscomb, Po-Lun Ma, Salil Mahajan, Mathew E. Maltrud, Azamat Mametjanov, Julie L. McClean, Renata B. McCoy, Richard B. Neale, Stephen F. Price, Yun Qian, Philip J. Rasch, J. E. Jack Reeves Eyre, William J. Riley, Todd D. Ringler, Andrew F. Roberts, Erika L. Roesler, Andrew G. Salinger, Zeshawn Shaheen, Xiaoying Shi, Balwinder Singh, Jinyun Tang, Mark A. Taylor, Peter E. Thornton, Adrian K. Turner, Milena Veneziani, Hui Wan, Hailong Wang, Shanlin Wang, Dean N. Williams, Phillip J. Wolfram, Patrick H. Worley, Shaocheng Xie, Yang Yang, Jin-Ho Yoon, Mark D. Zelinka, Charles S. Zender, Xubin Zeng, Chengzhu Zhang, Kai Zhang, Yuying Zhang, Xue Zheng, Tian Zhou, and Qing Zhu. The doe e3sm coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, 11(7):2089–2129, 2019.

[20] Benjamin Hillman, Peter Caldwell, Andrew G. Salinger, Luca Bertagna, Hassan Beydoun, Bogenschutz. Peter, Andrew Michael Bradley, Aaron Donahue, Christopher Eldred, James G. Foucar, Chris Golaz, Oksana Guba, Robert Jacob, Jeff Johnson, Noel Keen, Jayesh Krishna, Wuyin Lin, Weiran Liu, Kyle Pressel, Balwinder Singh, Andrew Steyer, Mark A. Taylor, Chris Terai, Paul Ullrich, Danqing Wu, and Xingqui Yuan. Scream: a performance-portable global cloud-resolving model based on the energy exascale earth system model. 6 2020.

[21] Sachin S. Wani and Sarita Thakar. Crank-nicolson type method for burgers equation. *International Journal of Applied Physics and Mathematics*, pages 324–328, 2013.

[22] Hwijae Son, Jin Woo Jang, Woo Jin Han, and Hyung Ju Hwang. Sobolev training for physics informed neural networks, 2021.

[23] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. *CoRR*, abs/2012.00926, 2020.

[24] Sapsis Themistoklis P. Output-weighted optimal sampling for bayesian regression and rare event statistics using few samples. *Proc. R. Soc.*, A. 476:20190834. 20190834, 2020.

[25] Di Qi and Andrew J Majda. Predicting extreme events for passive scalar turbulence in two-layer baroclinic flows through reduced-order stochastic models. *Communications in Mathematical Sciences*, 16(1):17–51, 2018.

[26] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-No el Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[27] A. B. M. Jeuken, P. C. Siegmund, L. C. Heijboer, J. Feichter, and L. Bengtsson. On the potential of assimilating meteorological analyses in a global climate model for the purpose of model validation. *Journal of Geophysical Research: Atmospheres*, 101(D12):16939–16950, 1996.

[28] Gabriel J. Kooperman, Michael S. Pritchard, Steven J. Ghan, Minghuai Wang, Richard C. J. Somerville, and Lynn M. Russell. Constraining the influence of natural variability to improve estimates of global aerosol indirect effects in a nudged version of the community atmosphere model 5. *Journal of Geophysical Research: Atmospheres*, 117(D23204), 2012.

[29] U. Lohmann and C. Hoose. Sensitivity studies of different aerosol indirect effects in mixed-phase clouds. *Atmospheric Chemistry and Physics*, 9(22):8917–8934, 2009.

[30] Guangxing Lin, Hui Wan, Kai Zhang, Yun Qian, and Steven J. Ghan. Can nudging be used to quantify model sensitivities in precipitation and cloud forcing? *Journal of Advances in Modeling Earth Systems*, 8(3):1073–1091, 2016.

[31] Oliver Watt-Meyer, Noah D. Brenowitz, Spencer K. Clark, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, and Christopher S. Bretherton. Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, 48(15):e2021GL092555, 2021.

[32] Christopher S. Bretherton, Brian Henn, Anna Kwa, Noah D Brenowitz, Oliver Watt-Meyer, Jeremy McGibbon, W. Andre Perkins, Spencer K. Clark, and Lucas Harris. Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Earth and Space Science Open Archive*, page 39, 2021.

[33] W. Lawrence Gates, James S. Boyle, Curt Covey, Clyde G. Dease, Charles M. Doutriaux, Robert S. Drach, Michael Fiorino, Peter J. Gleckler, Justin J. Hnilo, Susan M. Marlais, Thomas J. Phillips, Gerald L. Potter, Benjamin D. Santer, Kenneth R. Sperber, Karl E. Taylor, and Dean N. Williams. An overview of the results of the atmospheric model intercomparison project (amip i). *Bulletin of the American Meteorological Society*, 80(1):29–56, 1999.

[34] Richard W. Reynolds, Nick A. Rayner, Thomas M. Smith, Diane C. Stokes, and Wanqiu Wang. An improved in situ and satellite sst analysis for climate. *Journal of Climate*, 15(13):1609–1625, 2002.

[35] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[36] R. M. Hoesly, S. J. Smith, L. Feng, Z. Klimont, G. Janssens-Maenhout, T. Pitkanen, J. J. Seibert, L. Vu, R. J. Andres, R. M. Bolt, T. C. Bond, L. Dawidowski, N. Kholod, J.-I. Kurokawa, M. Li, L. Liu, Z. Lu, M. C. P. Moura, P. R. O'Rourke, and Q. Zhang. Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (ceds). *Geoscientific Model Development*, 11(1):369–408, 2018.

[37] L. Feng, S. J. Smith, C. Braun, M. Crippa, M. J. Gidden, R. Hoesly, Z. Klimont, M. van Marle, M. van den Berg, and G. R. van der Werf. The generation of gridded emissions data for cmip6. *Geoscientific Model Development*, 13(2):461–482, 2020.

[38] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.