DARPA ACTM Milestone 3 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)

May 1, 2022

1 Data set for Phase 1 (PNNL data)

Our aim is to construct an online bias correction method for climate modeling using machine learning (ML). The ML will predict a corrective tendency that can be applied to the prognostic state of the atmospheric model at each time step in order to reduce model biases. This method will be implemented in version 2 of the E3SM Atmospheric Model (EAMv2). EAMv2 solves the primitive equations on a cubed-sphere grid with horizontal discretization localized on individual elements. The lower-resolution configuration of EAMv2 is employed as the target model to be improved with the ML bias correction approach. Specifically, the low-resolution model employs approximately 1^o ($\sim$ 110 km) resolution in horizontal and 72 layers in vertical. The vertical layers extend from the Earth's surface to ~0.1 hPa (~64 km). The vertical grid spacing is uneven, with the layer height ranging from 20–100 m near the surface and up to 600 m near the model top. Our goal is to reduce the biases associated with the physics parameterizations with ML bias correction so that the coarse-grid EAMv2 produces similar climate statistics compared to reference data sets from an observationally-constrained reanalysis or the high-fidelity model.

In phase 1, the ML training data are constructed following a similar strategy described in Watt-Meyer et. al. (2021).[1] The "nudge-to-observations" approach is employed to estimate the biases in EAMv2 model state including temperature (T), humidity (Q), zonal wind (U) and meridional wind (V) for the ML training. Here, nudging constrains the model solution of X_m at every grid point toward the reference state of X_r by adding a linear relaxation term to the coarse EAMv2 model equation:

$$\frac{\partial \boldsymbol{X_m}}{\partial t} = \underbrace{\boldsymbol{D}\left(\boldsymbol{X_m}\right)}_{dynamics} + \underbrace{\boldsymbol{P}\left(\boldsymbol{X_m}\right)}_{physics} + \dot{\boldsymbol{R}} \tag{1}$$

$$\dot{\boldsymbol{R}} = -\underbrace{\frac{\boldsymbol{X_m} - \boldsymbol{X_r}}{\tau}}_{nudging} \tag{2}$$

where $\boldsymbol{X_m}$ and $\boldsymbol{X_r}$ refer to the state variables of U, V, T, Q from the EAMv2 predictions and the reference data sets, respectively. The first term on the right-hand side of Eq. 1 represents the effects of large-scale dynamics (e.g. large-scale advection etc.). The second term denotes the parameterized effects of physical processes such as clouds and convection that operate at scales smaller than the model grid and affect the overall dynamics of the system. The third term $\dot{\boldsymbol{R}}$ is a nudging tendency term that acts as an error correction for $\boldsymbol{X_m}$, calculated as the difference between $\boldsymbol{X_m}$ and $\boldsymbol{X_r}$, scaled by the relaxation time scale τ (Eq. 2).

Figure 1a shows the distribution of monthly mean zonal averaged temperature differences between the EAMv2 free-running simulations (i.e., CLIM) and ERA5 reanalysis (i.e. reference) in January 2010. Most model layers in the Tropics and mid-latitudes exhibit a cold temperature bias. In these regions, the positive temperature nudging tendencies in the nudged simulation act to correct the cold biases (Fig. 1b). Generally the time mean nudging tendency removes the systematic "background error" found in the EAMv2 free-running simulations. However, the nudging may not always help to reduce the systematic errors. For example, nudging both wind and temperature can produce a positive tendency of temperature in the northern hemisphere high-latitude (Fig. 1b), where the free-running simulations exhibit warm temperature biases, as shown in Fig. 1a, suggesting a role of positive feedback that amplifies the upper level temperature biases in the free-running simulations. Using a nudging strategy that constrains humidity in addition to wind and temperature produces rather different nudging tendencies (Fig. 1c), revealing the complex relationships between the nudging corrections and the state variables through the nonlinear governing equation (Eq. 1). Hence as discussed below, we design different nudging strategies to provide an ensemble of nudged simulations with different nudging tendencies and state variables for the ML training.

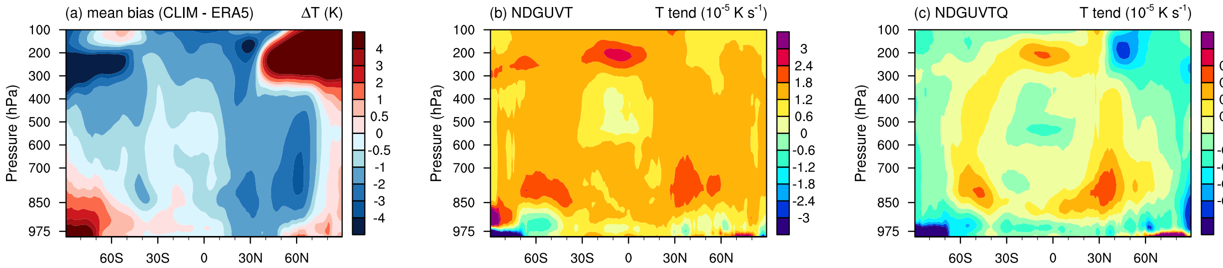


Figure 1: (a) monthly mean zonally averaged temperature differences (ΔT, unit: K) in January 2010 between ERA5 and EAMv2's free-running simulation (CLIM in Table 1), (b-c) monthly mean nudging tendencies of temperature (T tend, unit: K s^{-1}) from the simulation by nudging EAMv2 towards ERA5 reanalysis. The wind and temperature fields were nudged in the simulation (NDG_UVT_tau6 in Table 1) for panel (b), while the wind, temperature and humidity were nudged in the simulation (NDG_UVTQ_tau24 in Table 1) for panel (c). The y-axis of each panel shows the approximated pressure for the model levels in EAMv2.

For the "nudge-to-observations" approach in this project, the observations (i.e. reference data sets) are taken from the ERA5 reanalysis developed by the European Centre for Medium-Range Weather Forecasts (ECMWF).[2] The raw ERA5 reanalysis data are produced on a 0.25^o horizontal grid over the globe, which are spatially remapped to the cubed-sphere grid and the 72 model layers used by EAMv2, 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 EAMv2 and the reanalysis data are taken into account during the vertical interpolation. In addition, the 3-hourly ERA5 data for U, V, T, Q from December 2009 to November 2010 is selected as the reference data sets for the ML training to account for the seasonal and diurnal variations (see Group 1 in Table 1). As a first step, we use one year of data for training to reduce computational cost and archiving data storage, while temporal sampling of the training data will be extended as needed to achieve the best performance of the ML method.

The nudging tendency data for ML training are generated by performing EAMv2 simulations nudged towards the reference data sets (i.e. 3-hourly ERA5 reanalysis data). As shown in Eq. 2, the time-dependent nudging tendency is determined by the model ($\boldsymbol{X_m}$) and reference state ($\boldsymbol{X_r}$) as well as the empirical nudging relaxation time scale (τ). The nudging tendencies are calculated at every model step (30 minutes) following Eq. 2 in which $\boldsymbol{X_r}$ is obtained by linearly interpolating the 3-hourly ERA5 reanalysis data to the model time. Considering the uncertainties in the nudging tendencies arising from different combinations of the state variables in $\boldsymbol{X_r}$ (i.e. U, V, T, Q) to be nudged and from different choices of τ, we construct an ensemble of training data sets of $\dot{\boldsymbol{R}}$ (i.e. R_U , R_V , R_T , R_Q) by applying nudging:

- only to the horizontal winds with $\tau = 3$ (labeled "UV_tau3"), $\tau = 6$ (labeled "UV_tau6") and 24 hrs (labeled "UV_tau24"), respectively.
- to both winds and temperature with $\tau = 3$ (labeled "UVT_tau3"), 6 (labeled "UVT_tau6") and 24 hrs (labeled "UVT_tau24"), respectively.
- to winds, temperature, and humidity with $\tau = 3$ (labeled "UVTQ_tau3"), 6 (labeled "UVTQ_tau6") and 24 hrs (labeled "UVTQ_tau24"), respectively.

For each nudged simulation, we store the profiles of model state for every model grid column at two instances within a single model time step before and after the state variables are corrected by the nudging tendencies: "before nudging" (labeled as U_b, V_b, T_b, Q_b) and "after nudging" (labeled as U_a, V_a, T_a, Q_a). We also store the averaged nudging tendencies (R_U, R_V, R_T, R_Q) every three hours for every model grid column. Table 1 summarizes the information for each nudged simulation (Group 2). In addition, all nudged simulations are run from October 2009 to November 2010. The first two months are treated as model spin up with nudging, and the remaining data from December 2009 to November 2010 are used to construct the 1-year ML training data. We plan to run additional nudged EAMv2 simulations with different nudging strategies (e.g., nudging relative humidity instead of specific humidity) to examine if creating a larger ensemble of training data capturing more diverse nudging strategies can help improve the performance of ML method.

In this project, the 3-hourly averaged instead of instantaneous nudging tendencies are used for the ML training following the recommendation of Watt-Meyer et al. (2021).[1] Watt-Meyer et al. (2021)[1] showed that instantaneous nudging tendency profiles can have complex vertical structure that vary greatly in space and time and degrade the ML skill, while time-mean nudging tendencies are large and well-learned by the ML. In this way, our ML scheme predicts vertical profiles of 3-hour averaged nudging tendencies in each EAMv2 grid column. During the training period, the ML learns the mapping between the "before nudging" model state at time t and the mean bias correction tendency $\dot{\boldsymbol{M}}$ from t to $t + \Delta t$:

$$\dot{\boldsymbol{M}}(t, t + \Delta t) = \underbrace{\boldsymbol{G}[U_b(t), V_b(t), T_b(t), Q_b(t)]}_{ML\ model} \tag{3}$$

Where Δt is 3 hours in our case, and $\dot{\boldsymbol{M}}(t, t + \Delta t)$ is the mean bias correction tendency averaged from t to $t + \Delta t$ predicted by the ML, and $\boldsymbol{G}$ denotes the mapping function derived by the ML algorithm. We expect that $\dot{\boldsymbol{M}}(t, t+\Delta t)$ evolves similarly to the 3-hourly averaged nudging tendencies $\dot{\boldsymbol{R}}$ if our ML approach is successful. The "after nudging" model state (i.e. U_a, V_a, T_a, Q_a) for offline ML application, and the model state(i.e. U, V, T, Q) from a coarse-grid EAMv2 free-running simulation (i.e. CLIM) with online ML application will be used for the verification and validation of the ML approach, which will be further discussed in Section 3.

Table 1: List of training data for machine learning. Note: nudging is applied at every model physics time step (0.5-hr) for EAMv2. The variables R denote the nudging tendencies terms for the corresponding model state variable. The data is sampled every 3 hours during December 2009 - November 2010 at each grid point over the globe. The model state variable is the instantaneous model output, while the nudging tendencies are averaged values during a 3-hr period for each time sample. The data on the original EAMv2 model grid (cubed-sphere grid) are available at `https://portal.nersc.gov/cfs/e3sm/zhan391/darpa_temporary_data_share/SE_PG2/`, while the regridded data on the lat-lon grid with a 1^o resolution are available at `https://portal.nersc.gov/cfs/e3sm/zhan391/darpa_temporary_data_share/FV_180x360`. The detailed description for the datasets can be found in the "READ_ME.TXT" file in the above directory.

Group	Short name.	Nudged variables	Relaxation scale (τ)	Training/verification data
1	Reference (ERA5)	None	N/A	U, V, T, Q
2	NDG_UV_tau3	U, V	3 hr	R_U, R_V , U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UV_tau6	U, V	6 hr	R_U, R_V , U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UV_tau24	U, V	24 hr	R_U, R_V , U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVT_tau3	U, V T	3 hr	R_U, R_V, R_T, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVT_tau6	U, V T	6 hr	R_U, R_V, R_T, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVT_tau24	U, V T	24 hr	R_U, R_V, R_T, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVTQ_tau3	U, V T, Q	3 hr	R_U, R_V, R_T, R_Q, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVTQ_tau6	U, V T, Q	6 hr	R_U, R_V, R_T, R_Q, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
2	NDG_UVTQ_tau24	U, V T, Q	24 hr	R_U, R_V, R_T, R_Q, U_b, V_b, T_b, Q_b, U_a, V_a, T_a, Q_a
3	CLIM	None	N/A	U, V, T, Q

2 Data set for beta testing

The purpose of beta testing is to implement and fine-tune the proposed methodology on a simple climate model. This model will act as a computationally cheap surrogate for the final E3SM dataset. As a result, a multitude of ideas can be implemented fast and cheaply. Hence, the resulting methodology after this stage will be more reliable to succeed on the more complicated E3SM model. To this end, a two-layer quasi-geostrophic (QG)[3] model is employed to mimic the velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$ of atmospheric flows. The turbulent flow is described by the dimensionless evolution equation

$$\frac{\partial q_j}{\partial t} + \boldsymbol{v}_j \cdot \boldsymbol{\nabla} q_j + \left(\beta + k_d^2 U_j\right) \frac{\partial \psi_j}{\partial x} = -\delta_{2,j} r \Delta \psi_j - \nu \Delta^s q_j, \tag{4}$$

where $j = 1, 2$ corresponds to the upper and lower layer respectively and the flow is defined in the horizontal domain $(x, y) \in [0, 2\pi]^2$. Doubly periodic boundary conditions are assumed. 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. 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

$$q_j = \Delta \psi_j + \frac{k_d^2}{2} \left(\psi_{3-j} - \psi_j\right), \quad j = 1, 2. \tag{5}$$

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.

The quasigeostrophic model described by eq. (4) is used here to simulate mid latitude and high latitude atmospheric flows that are forced by an imposed shear current.[4] The principal parameter to be varied is the beta-plane approximation parameter β. To locate an interval of values for β that correspond to physically relevant simulations, the Coriolis acceleration need to be taken into account. Taking into account the dimensionless form of eq. (4), the Coriolis frequency at some latitude ϕ_0 is defined as

$$f = 2 \frac{\Omega L_y}{U_{\text{scale}}} \sin\left(\phi_0 + L_y \frac{y}{R}\right), \tag{6}$$

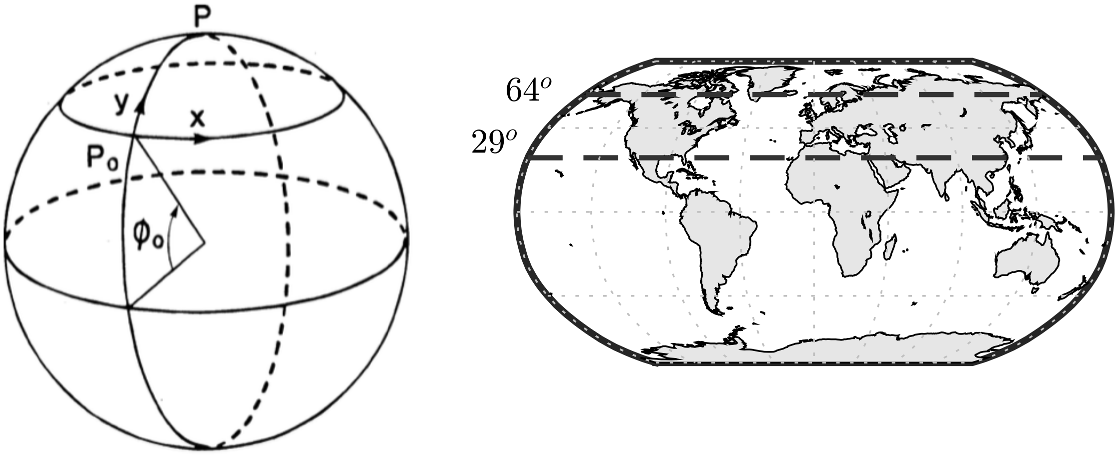
where Ω is the frequency of the rotation of the earth and set to $\Omega = 7.2925 \cdot 10^{-5}$ [rad/sec]. R is the radius of the earth and set to $R = 6378$ [km]. Parameters L_y and U_{scale} are scales for the meridional extent and velocity field respectively. A schematic depicting the domain on a globe is seen at fig. 2, subfigure (a). Utilizing a Taylor expansion, the beta-plane approximation coefficient arises from

$$\begin{aligned} f &= 2 \frac{\Omega L_y}{U_{\text{scale}}} \sin \phi_0 + 2\Omega \frac{L_y^2}{U_{\text{scale}}} \frac{\cos \phi_0}{R} y + O\left(L_y^2 \frac{y^2}{R^2}\right) \\ &\approx 2 \frac{\Omega L_y}{U_{\text{scale}}} \sin \phi_0 + \underbrace{2\Omega \frac{L_y^2}{U_{\text{scale}}} \frac{\cos \phi_0}{R}}_{\beta} y \end{aligned} . \tag{7}$$

For atmospheric flows, a standard assumption is $L_y \sim 1000\,[\text{km}]$ and $U_{\text{scale}} \sim 10\,[\text{m/sec}]$. Hence, are range of β values capable of include midlatitude and high latitude cases, correspond to $\beta \in [1, 2]$. Indeed, one can check that $\beta = 1$ corresponds to $\phi_0 \approx 29^{\text{o}}$ and $\beta = 2$ corresponds to $\phi_0 \approx 64^{\text{o}}$. The extent of this regime on a globe map is seen in fig. 2, subfigure (b). Finally, table 2 contains parameter values for a mid latitude and a high latitude regime.

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

Table 2: Parameter values for different atmosphere regimes.



(a)

(b)

Figure 2: (a) Schematic of beta-plane approximation on a globe. (b) The meridional extent that the simulations with $\beta \in [1, 2]$ correspond to.

For training purposes, an appropriate nudged data set needs to be determined first. Nudging is enforced due to chaotic divergence in turbulent flows. This implies that free running coarse-scale data will very quickly diverge from their fine-scale conuterpart, despite having the same flow parameters and initial conditions. To that end, to produce coarse-scale simulations for training, a relaxation term Q is added to the evolution eq. (4) to become

$$\frac{\partial q_j}{\partial t} + \boldsymbol{v}_j \cdot \boldsymbol{\nabla} q_j + \left(\beta + k_d^2 U_j\right) \frac{\partial \psi_j}{\partial x} = -\delta_{2,j} r \Delta \psi_j - \nu \Delta^s q_j - \frac{1}{\tau}\left(q_j - \mathcal{H}\left[q_j^{\text{ref}}\right]\right), \qquad (8)$$

where $j = 1, 2$. Parameter τ is a relaxation timescale to be determined, and $\mathcal{H}$ is an operator that maps q_j^{ref} to the coarse resolution. As discussed in Milestone 2, τ is set to $\tau = 16$. This value allows for a simulation that follows the reference data while having statistics close to that of the coarse-scale simulation.

To remedy the energy spectra differences between the testing data $(\psi_1^{\text{coarse}}, \psi_2^{\text{coarse}})$ and training data $\left(\psi_1^{\text{nudge}}, \psi_2^{\text{nudge}}\right)$, a new method is developed and employed. The process is called ‘Reverse Spectral Nudging’ with its purpose being to match the energy spectrum of the nudged solution to that of the coarse-scale solution to improve the training process. Hence, while traditional nudging schemes correct the coarse-scale solution with data from the reference solution, the proposed scheme further processes the nudged data by matching its energy spectrum to that of the corresponding free running coarse-scale flow. The corrected nudged data is termed as $\left(\psi_1^{\text{RS-nudge}}, \psi_2^{\text{RS-nudge}}\right)$ and defined as

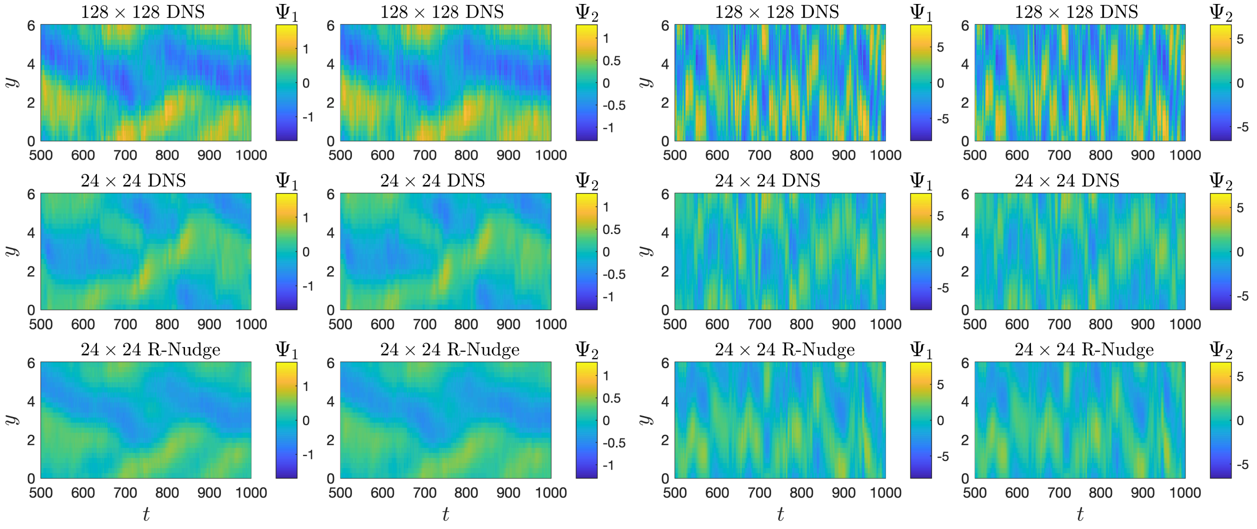
$$\psi_i^{\text{RS-nudge}}(x,y,t) = \sum_{k,l} R_{k,l}\hat{\psi}_{k,l}^{\text{nudge}}(t)e^{i(kx+ly)},, \tag{9}$$

where $\psi_{k,l}^{\text{nudge}}(t)$ are the spatial Fourier coefficients of $\left(\psi_1^{\text{nudge}}, \psi_2^{\text{nudge}}\right)$ and

$$R_{k,l} = \sqrt{\frac{\mathcal{E}_{k,l}^{\text{coarse}}}{\mathcal{E}_{k,l}^{\text{nudge}}}}, \quad \text{and} \quad \mathcal{E}_{k,l} = \frac{1}{T}\int_0^T \hat{E}_{k,l}(t)\mathrm{d}t = \frac{1}{T}\int_0^T |\hat{\psi}_{k,l}(t)|^2\mathrm{d}t. \tag{10}$$

An important property of this scheme is that the new data have exactly the energy spectrum of the free running coarse simulation, meaning that the training and testing data come from the same distributions. This property improves significantly the accuracy of the resulted ML scheme.

For the current numerical investigation, the reference horizontal resolution is set to 128 × 128. The coarse-scale simulations are set to 24 × 24, a resolution that is inadequate to correctly capture the statistics of such flows. Both sets of simulations have the same temporal resolution. The flow is solved using spectral method and the time evolution scheme is the explicit 4th-order Runge-Kutta.



(a)

(b)

Figure 3: Predictions of zonally-averaged streamfunctions for reference simulation, coarse-scale free runing simulation and coarse-scale R-nudged simulation with (a) $\beta = 2.0$ and (b) $\beta = 1.0$.

Despite the simple structure of the model, quasi-geostrophic flows experience intermittent energy cascades. This phenomenon renders the effects of small scales to large-scale flow features crucial. These energy interactions yield non-Gaussian statistics that significantly change as parameters vary. Hence, despite the small interval β is allowed to vary within, a multitude of different statistical behaviours can be observed. As can be seen in fig. 3, different values of β result in flows with drastically different timescales for the flow dynamics.

3 Metrics

3.1 Metrics for verification and validation of offline ML

The 1-year data listed in Table 1 is used for offline training and testing of the DeepONet ML model in our project. A procedure is designed to (a) randomly select a subset of temporal samples within the 1-year period as the ML training data from across the diurnal and seasonal cycles, and (b) test the offline model skill using the remaining data that is independent from the training data. The performance of the offline ML model, which we call offline skill, is tested in two ways.

First, metrics are used to determine how well the bias correction tendency $\dot{\boldsymbol{M}}$ predicted by the ML model using the before nudged state variables as predictors resembles the nudging tendency $\dot{\boldsymbol{R}}$ from the nudged simulations. This is evaluated by directly comparing $\dot{\boldsymbol{M}}$ from ML with $\dot{\boldsymbol{R}}$ in the training data sets, excluding the samples used in training the ML. Both $\dot{\boldsymbol{M}}$ and $\dot{\boldsymbol{R}}$ are vectors containing four state variables (i.e. U, V, T, Q, see Table 3 for a detailed description). The following metrics are defined:

$$\begin{aligned}
\overline{\dot{\boldsymbol{R}}}(x,y,z) =& \frac{1}{N_t}\sum_{t=1}^{N_t}\Big(\dot{\boldsymbol{R}}(x,y,z,t)\Big)\\
\boldsymbol{S}_r(x,y,z) =& \sum_{t=1}^{N_t}\Big(\dot{\boldsymbol{M}}(x,y,z,t)-\dot{\boldsymbol{R}}(x,y,z,t)\Big)^2\\
\boldsymbol{S}_t(x,y,z) =& \sum_{t=1}^{N_t}\Big(\dot{\boldsymbol{R}}(x,y,z,t)-\overline{\dot{\boldsymbol{R}}}(x,y,z)\Big)^2\\
\boldsymbol{S}^2(x,y,z) =& 1-\frac{\boldsymbol{S}_r(x,y,z)}{\boldsymbol{S}_t(x,y,z)}
\end{aligned} \tag{11}$$

Where $\boldsymbol{S}_r$ represents the residuals sum of squares, and $\boldsymbol{S}_t$ represents the total sum of squares (proportional to the variance of $\dot{\boldsymbol{R}}$) over the time dimension. $\boldsymbol{S}^2$ refers to the coefficient of determination, which is used to evaluate the similarity between $\dot{\boldsymbol{M}}$ and $\dot{\boldsymbol{R}}$:

- The best performance corresponds to $\boldsymbol{S}_r = 0$ and $\boldsymbol{S}^2 = 1$.
- If the ML-predicted correction tendency $\dot{\boldsymbol{M}}$ always resembles $\overline{\dot{\boldsymbol{R}}}$, we will have $\boldsymbol{S}^2 = 0$.
- If the ML-predicted correction tendency does not resemble the nudging tendency $\dot{\boldsymbol{R}}$ (i.e., the differences between $\dot{\boldsymbol{M}}$ and $\dot{\boldsymbol{R}}$ are larger than those between $\dot{\boldsymbol{R}}$ and $\overline{\dot{\boldsymbol{R}}}$), then $\boldsymbol{S}^2$ is negative.

Table 3: Description of notation. Notes: the (x, y, z, t) is corresponding to the (latitude, logitude, levels, time) dimension in the EAMv2 model output. Each notation contains the four state variables (i.e. U, V, T, Q) that are interested in this project

Notation	Dimension	Source	Description
$\dot{\boldsymbol{M}}$	(x, y, z, t)	ML output	3-hourly averaged correction tendency from ML
$\dot{\boldsymbol{R}}$	(x, y, z, t)	Training data	3-hourly averaged nudging tendency
$\boldsymbol{X_r}$	(x, y, z, t)	Training data	3-hourly state variable from reference data set
$\boldsymbol{X_m}$	(x, y, z, t)	Post-processing	3-hourly state variable derived from ML
$\boldsymbol{X_n}$	(x, y, z, t)	Post-processing	3-hourly state variable derived from nudging
$\boldsymbol{X_b}$	(x, y, z, t)	Training data	3-hourly state variable at "before nudging" location
$\boldsymbol{X_a}$	(x, y, z, t)	Training data	3-hourly state variable at "after nudging" location

Second, the offline skill of the ML model is also evaluated by how much the ML bias correction improves the skill of the model state simulated by EAMv2. Eq. 2 describes the relationship among the nudging tendency, model prediction and the reference state. As the nudging tendency (i.e. $\dot{\boldsymbol{R}}$) and the corresponding model predicted state (referred to as "$\boldsymbol{X_b}$") are known, we can estimate the "reference" state reached by EAMv2 (referred to as "$\boldsymbol{X_n}$") with:

$$\boldsymbol{X_n}(x,y,z,t) = \boldsymbol{X_b}(x,y,z,t) + \dot{\boldsymbol{R}}(x,y,z,t)\,\tau \tag{12}$$

where τ is the nudging relaxation time scale and $\dot{\boldsymbol{R}}$ is the 3-hour mean nudging tendency. Table 3 lists detailed descriptions of other notation.

Similarly, the "reference" state estimated with $\dot{\boldsymbol{M}}$ (referred to as "$\boldsymbol{X_m}$") can be written as:

$$\boldsymbol{X_m}(x,y,z,t) = \boldsymbol{X_b}(x,y,z,t) + \dot{\boldsymbol{M}}(x,y,z,t)\,\tau \tag{13}$$

In this way, $\boldsymbol{X_n}(x,y,z,t)$ and $\boldsymbol{X_m}(x,y,z,t)$ can be directly compared and evaluated with the true reference state $\boldsymbol{X_r}(x,y,z,t)$ from the reference data sets listed in Table 1. In addition to the commonly used statistical measures including mean biases and root-mean-square difference (RMSD), we also introduce two metrics based on anomaly correlation (AC) to quantify the consistencies of temporal variation and the large-scale spatial pattern between the $\boldsymbol{X_m}(x,y,z,t)$ (or $\boldsymbol{X_n}(x,y,z,t)$) and $\boldsymbol{X_r}(x,y,z,t)$:

- **Temporal Correlation Coefficient (TCC)**, which provides a measure of the consistency of time evolution of anomaly fields between the EAMv2 simulations and the observations at each grid points.

$$TCC(x,y,z) = \frac{\sum_{t=1}^{N_t}\left(\boldsymbol{X'_m}(x,y,z,t) - \overline{\boldsymbol{X'_m}}(z)\right)\left(\boldsymbol{X'_r}(x,y,z,t) - \overline{\boldsymbol{X'_r}}(z)\right)}{\sqrt{\sum_{t=1}^{N_t}\left(\boldsymbol{X'_m}(x,y,z,t) - \overline{\boldsymbol{X'_m}}(z)\right)^2}\sqrt{\sum_{t=1}^{N_t}\left(\boldsymbol{X'_r}(x,y,z,t) - \overline{\boldsymbol{X'_r}}(z)\right)^2}} \tag{14}$$

- **Spatial Anomaly Correlation Coefficient (ACC)**, which provides a measure of consistency of spatial patterns of weather systems between the EAMv2 simulations and the observations at a forecast time. It is the spatial correlation between a forecast anomaly from climatology and a verifying analysis anomaly from climatology.

$$ACC(z,t) = \frac{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y} w(x,y)\left(\boldsymbol{X'_m}(x,y,z,t) - \overline{\boldsymbol{X_m}}(z)\right)\left(\boldsymbol{X'_r}(x,y,z,t) - \overline{\boldsymbol{X_r}}(z)\right)}{\sqrt{\sum_{x=1}^{N_x}\sum_{x=1}^{N_y} w(x,y)\left(\boldsymbol{X'_m}(x,y,z,t) - \overline{\boldsymbol{X_m}}(z)\right)^2}\sqrt{\sum_{x=1}^{N_x}\sum_{x=1}^{N_y} w(x,y)\left(\boldsymbol{X'_r}(x,y,z,t) - \overline{\boldsymbol{X_r}}(z)\right)^2}} \tag{15}$$

The anomaly terms and location weights (w) in equations 14 and 15 are defined as follows.

$$\begin{aligned}
\boldsymbol{X'_m}(x,y,z,t) &= \boldsymbol{X_m}(x,y,z,t) - \frac{1}{N_t}\sum_{t=1}^{N_t}\boldsymbol{X_m}(x,y,z,t) \\
\boldsymbol{X'_r}(x,y,z,t) &= \boldsymbol{X_r}(x,y,z,t) - \frac{1}{N_t}\sum_{t=1}^{N_t}\boldsymbol{X_r}(x,y,z,t) \\
\overline{\boldsymbol{X'_m}}(z) &= \frac{1}{\overline{w}}\sum_{t=1}^{N_t}\sum_{x=1}^{N_x}\sum_{x=1}^{N_y} w(x,y)\boldsymbol{X'_m}(x,y,z,t) \\
\overline{\boldsymbol{X'_r}}(z) &= \frac{1}{\overline{w}}\sum_{t=1}^{N_t}\sum_{x=1}^{N_x}\sum_{x=1}^{N_y} w(x,y)\boldsymbol{X'_r}(x,y,z,t)
\end{aligned} \tag{16}$$

$$\overline{w} = \sum_{x=1}^{N_x}\sum_{x=1}^{N_y} w(x,y);\;\; w(x,y) = \cos\left(\text{latitude}(x,y)\right) \tag{17}$$

Table 4 lists other metrics that are used in our evaluation, which are also widely used in the climate community to evaluate the climate model fidelity in terms of climate statistics. The evaluation with these metrics will help us assess if the ML corrections improve the skills of the large-scale dynamics and thermodynamics simulated by EAMv2. We should point out that our ML model is trained “offline” to test different methods in an efficient way. The state variable $\boldsymbol{X_m}$ derived from the ML model does not consider the impacts of the ML bias correction on other parts of the EAMv2 model. Hence the evaluation described so far focuses only on evaluating the ML model for postprocessing of climate model output, which is different from its skill for online bias correction of climate simulations to be described in the next section.

3.2 Metrics for verification and validation of online ML

Most of the metrics for“offline skill” can also be used to assess the “online” ML model performance, but the evaluation will focus on separate simulations with a longer time period for climate science. A set of the EAMv2 simulations with the ML bias correction model using the same configuration as the “CLIM” in Table 1 and Section 1 will be conducted for evaluation. The metrics of climate statistics as mentioned in previous section and summarized

Table 4: Description of metrics

Metrics	Description
Global mean	Average over globe,
Regional mean	Average over an interesting region
Spatial distribution	Global or regional mapping plot to show the patterns of an interested quantity
Vertical profiles	Profilings to indicate the vertical variation of the interested quantity
Zonal mean	Average over all longitude points for the interested quantity
Meridional mean	Average over all latitude points the interested quantity
Hovmöller Diagram	Time-latitude or time-longitude variation of the zonal or meridional mean quantities

in Table 4 will be calculated with respect to the observations and reanalysis products (e.g ERA5 reanalysis) to evaluate whether our ML model improves the model fidelity of EAMv2. In addition to the state variables, which are directly bias corrected by the ML model, we will also extend the metrics to other quantities including the precipitation, radiative fluxes, and cloud-related quantities, which are simulated by the physics parameterizations of EAMv2. These quantities are of primary interest to climate scientists and we expect that improvements in simulating the model states through the ML bias-correction would also improve the simulations of these fluxes.

The ultimate goal of our project is to improve the low-resolution EAMv2 model fidelity with a better simulation of the large-scale atmospheric circulation that can strongly affect extreme weather and climate events. One way to evaluate the large-scale atmospheric circulation is to look for the representation of recurrent spatio-temporal patterns, commonly referred to as weather regimes in the climate model simulations. In this project, we will employ empirical orthogonal function (EOF) and clustering analysis which are widely used in the climate community[5–7] to construct metrics. These metrics will be used to assess whether the ML bias-corrected EAMv2 produces a better representation of the weather regimes than the free-running "CLIM" simulation.

Finally, when the ML model is applied "online", the feedbacks of ML bias correction on other parts of the EAMv2 model will be included, which could lead to climate drifts or model instability. Thus, we will design extra metrics to quantitatively evaluate climate drifts and model stability in EAMv2 with the ML model developed by our project.

3.3 Metrics for beta testing

For the implementation of the methodology on the QG problem, metrics need to be determined both for the training and testing phase. The difference between the two processes is highlighted because they are carried out with different input data. Training is carried out with R-nudged data, while testing is carried out with free-running coarse scale simulations. As a result, during training metrics that showcase the success of the training process are of importance. A group of such metrics are related to the hybrid's scheme ability to correct phase errors between the R-nudged and reference data. The metrics are the following:

(i) Evolution of bias of zonally averaged streamfunctions:

$$\begin{aligned} \delta\Psi_i^{\mathrm{ML}}(y) &= \frac{1}{L_x}\int_0^{L_x} \psi_i^{\mathrm{ref}}(x,y)\,\mathrm{d}x - \frac{1}{L_x}\int_0^{L_x} \psi_i^{\mathrm{ML}}((x,y)\,\mathrm{d}x \\ &= \Psi_i^{\mathrm{ML}}(y) - \Psi_i^{\mathrm{ref}}(y). \end{aligned} \tag{18}$$

(ii) Temporal Correlation Coefficient, for a region $(x,y) \in [x_1,x_2]\times[y_1,y_2]$ and a time-interval $t \in [t_0, t_0+T]$:

$$\mathrm{TCC}\left(\psi_i; t_0\right) = \frac{1}{\Delta x \Delta y}\int_{x_1}^{x_2}\int_{y_1}^{y_2} \frac{\int_{t_0}^{t_0+T}\left(\psi_i^{\mathrm{ML}} - \overline{\psi_i^{\mathrm{ML}}}\right)\left(\psi_i^{\mathrm{ref}} - \overline{\psi_i^{\mathrm{ref}}}\right)\mathrm{d}t}{\sqrt{\int_{t_0}^{t_0+T}\left(\psi_i^{\mathrm{ML}} - \overline{\psi_i^{\mathrm{ML}}}\right)^2\mathrm{d}t\int_{t_0}^{t_0+T}\left(\psi_i^{\mathrm{ref}} - \overline{\psi_i^{\mathrm{ref}}}\right)^2\mathrm{d}t}}\mathrm{d}x\mathrm{d}y, \tag{19}$$

where $\Delta x = x_2 - x_1$, $\Delta y = y_2 - y_1$ and

$$\overline{\psi_i} = \frac{1}{\Delta x \Delta y T}\int_{x_1}^{x_2}\int_{y_1}^{y_2}\int_{t_0}^{t_0+T} \psi_i \mathrm{d}x\mathrm{d}y\mathrm{d}t. \tag{20}$$

(iii) Spatial Anomaly Correlation Coefficient, for a region $(x,y) \in [x_1,x_2]\times[y_1,y_2]$ and a time-interval $t \in [t_0, t_0+T]$:

$$\mathrm{ACC}\left(\psi_i; t_0\right) = \frac{1}{T}\int_{t_0}^{t_0+T} \frac{\int_0^{L_x}\int_0^{L_y}\left(\psi_i^{\mathrm{ML}} - \overline{\psi_i^{\mathrm{ML}}}\right)\left(\psi_i^{\mathrm{ref}} - \overline{\psi_i^{\mathrm{ref}}}\right)\mathrm{d}x\mathrm{d}y}{\sqrt{\int_0^{L_x}\int_0^{L_y}\left(\psi_i^{\mathrm{ML}} - \overline{\psi_i^{\mathrm{ML}}}\right)^2\mathrm{d}x\mathrm{d}y\int_0^{L_x}\int_0^{L_y}\left(\psi_i^{\mathrm{ref}} - \overline{\psi_i^{\mathrm{ref}}}\right)^2\mathrm{d}x\mathrm{d}y}}\mathrm{d}t. \tag{21}$$

As for testing, this project focuses on correctly predicting the statistical quantities of the simulated QG flows. Of particular interest are the tails of probability density functions (PDFs) of the streamfunctions as β varies.

Furthermore, PDFs of the amplitude of large-scale wavenumbers are of interest. These include $\hat{\psi}_{(1,0)}, \hat{\psi}_{(0,1)}$ and $\hat{\psi}_{(1,1)}$. The logarithmic difference between the reference PDFs and the ones predicted by the hybrid scheme can be computed as

$$\epsilon_f = \int_{\theta\in\Theta}\left(\log_{10} f^{\mathrm{ref}} - \log_{10} f^{\mathrm{ML}}\right)\mathrm{d}\theta, \tag{22}$$

where f corresponds to the PDF of a quantity of interest. Finally, energy spectra will also be compared between the reference solution and the hybrid scheme. This metric is given by the formula

$$\begin{aligned} E &= \sum_{k,l}\left(\log_{10}\hat{E}_{k,l}^{\mathrm{ref}} - \log_{10}\hat{E}_{k,l}^{\mathrm{ML}}\right) \\ &= \sum_{k,l}\left(\log_{10}\left[(k^2+l^2)\left|\hat{\psi}_{k,l}^{\mathrm{ref}}\right|^2\right] - \log_{10}\left[(k^2+l^2)\left|\hat{\psi}_{k,l}^{\mathrm{ML}}\right|^2\right]\right). \end{aligned} \tag{23}$$

References

[1] 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.

[2] 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.

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

[4] 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.

[5] S. V. Singh and R. H. Kripalani. Application of extended empirical orthogonal function analysis to interrelationships and sequential evolution of monsoon fields. *Monthly Weather Review*, 114(8):1603 – 1611, 1986.

[6] Neeti Neeti and J. Ronald Eastman. Novel approaches in extended principal component analysis to compare spatio-temporal patterns among multiple image time series. *Remote Sensing of Environment*, 148:84–96, 2014.

[7] F Fabiano, HM Christensen, K Strommen, P Athanasiadis, A Baker, R Schiemann, and S Corti. Euro-atlantic weather regimes in the primavera coupled climate simulations: impact of resolution and mean state biases on model performance. *Climate Dynamics*, 54(11):5031–5048, 2020.