Improving dynamical climate predictions with machine learning: insights from a twin experiment framework

Zikang He[1,2,3], Julien Brajard[3], Yiguo Wang[3], Xidong Wang[1,2], and Zheqi Shen[1,2]

[1]Key Laboratory of Marine Hazards Forecasting, Ministry of Natural Resources, Hohai University, Nanjing, 210098, China

[2]College of Oceanography, Hohai University, Nanjing, 210098, China

[3]Nansen Environmental and Remote Sensing Center, Bergen, N-5007, Norway

Correspondence: Xidong Wang (xidong_wang@hhu.edu.cn)

Abstract. Systematic errors in dynamical climate models remain a significant challenge to accurate climate predictions, particularly when modeling the nonlinear coupling between the atmosphere and ~~oceans~~ocean. Despite notable advances in dynamical climate modeling that have improved our understanding of climate variability, these systematic errors can still degrade ~~predictive~~ prediction skills. In this study, we adopt a twin experiment framework with a reduced-order coupled atmosphere-ocean model to explore the utility of machine learning in mitigating these errors. Specifically, we train a data-driven model on data assimilation increments to learn and emulate the underlying dynamical climate model error, which is then integrated with the dynamical climate model to form a hybrid ~~system~~model. Comparison experiments show that the hybrid model consistently outperforms the standalone dynamical climate model in predicting atmospheric and oceanic variables. Further investigation using hybrid models that correct only atmospheric or only oceanic errors reveals that atmospheric corrections are essential for improving short-term ~~forecasts~~predictions, while concurrently addressing both atmospheric and oceanic errors yields superior performance in long-term climate prediction.

1 Introduction

Climate prediction aims at predicting the future state of the climate system based on the initial conditions and external forcings (e.g., greenhouse gases and aerosols) covering various lead times from seasons to decades (Merryfield et al., 2020). It helps scientists, policymakers, and communities in understanding potential risks and impacts. It differs from climate projections that focus primarily on capturing long-term climate trends and patterns from several decades to centuries by anticipating changes in external forcings and their impact on the climate system.

Dynamical climate models, such as atmosphere-ocean coupled general circulation models, have been widely used for climate predictions (e.g., Doblas-Reyes et al., 2013b; Boer et al., 2016). Uncertainties in initial conditions fed to dynamical climate models and model errors are two critical sources that limit the prediction skill of dynamical climate models. To reduce the uncertainties of initial conditions, climate prediction centers (Balmaseda and Anderson, 2009; Doblas-Reyes et al., 2013a) have been evolving towards the use of data assimilation (DA, Carrassi et al., 2018) which combines observations with the dynamical climate models to estimate ~~best the state~~ the best initial conditions of the climate ~~system~~prediction (Penny and Hamill, 2017). Model errors can arise from a variety of sources, including model parameterizations (Palmer, 2001), unresolved physical processes (Moufouma-Okia and Jones, 2015), and numerical approximations (Williamson et al., 1992). Despite substantial efforts to improve dynamical climate models, these errors remain notably large (e.g., Richter, 2015; Palmer and Stevens, 2019; Richter and Tokinaga, 2020; Tian and Dong, 2020).

There is a growing interest in utilizing machine learning (ML) techniques to address errors in ~~the dynamical~~ a dynamical climate model. ML can be employed to construct a data-driven predictor of model errors, which can then be integrated with the dynamical climate model to create a hybrid statistical-dynamical model (e.g., Watson, 2019; Farchi et al., 2021a; Brajard et al., 2021; Watt-Meyer et al., 2021; Bretherton et al., 2022; Chen et al., 2022; Gregory et al., 2024).

Some notable studies ~~(e.g., Watson, 2019; Farchi et al., 2021a) have~~ (e.g., Watson, 2019; Brajard et al., 2021; Farchi et al., 2021a, 2023) focused on methodological developments within low-order or simplified coupled models operating in an idealized framework where the ground truth is known. For example, Farchi et al. (2021a) investigated two approaches in a two-scale Lorenz model, both of which are potential candidates for implementation in operational systems. One approach involves correcting the so-called resolvent of the dynamical climate model (i.e., modifying the model output after each numerical integration of the model). The other approach entails adjusting the ordinary or partial differential equation governing the model tendency before the numerical integration of the model. Similarly, Watson (2019) examined the tendency correction approach in the Lorenz 96 model. Brajard et al. (2021) explored the resolvent correction approach in the two-scale Lorenz model as well as in a low-order coupled atmosphere-ocean model called the Modular Arbitrary-Order Ocean-Atmosphere Model (MAOOAM, De Cruz et al., 2016). Their study aimed to infer the model errors associated with unresolved processes within the dynamical ~~model. In these works, the hybrid model is tested in an idealized setting in which initial conditionsare perfectly known. In realistic climate predictions, uncertainty in initial conditions is generally represented as an ensemble of initial conditions, and an ensemble of predictions is obtained (Wang et al., 2019). To our knowledge,~~ climate model. While Brajard et al. (2021) conducted prediction experiments using perfect initial conditions, more recent studies such as Farchi et al. (2023) examined the performance of hybrid models ~~under imperfect initial conditions—particularly when using an ensemble of forecasts—has not been thoroughly assessed. Moreover, it remains unclear which component of a coupled system contributes the most critical model error to climate predictions~~initialized with imperfect conditions, using a two-layer quasi-geostrophic (QG) model. Despite recent efforts to incorporate more realistic settings, hybrid models are still frequently evaluated under idealized conditions in which the initial state, taken from the same model as the reference, is assumed to be perfectly known.

Several other investigations ~~(e.g., Bonavita and Laloyaux, 2020; Watt-Meyer et al., 2021; Bretherton et al., 2022; Chen et al., 2022) have~~ tested ML-based error correction methods in realistic ~~weather or climate models. However, in the real framework, the ground truth is unknown and the error characteristics are complex. Moreover, observation for training, validation, and testing is relatively limited.~~

~~In this study, we aim to utilize the low-order coupled atmosphere-ocean model MAOOAM (section 2) to investigate the potential~~ numerical weather prediction (NWP) (e.g., Bonavita and Laloyaux, 2020; Watt-Meyer et al., 2021; Bretherton et al., 2022; Chen e . Bonavita and Laloyaux (2020) demonstrated that ML can emulate model error corrections derived from weak-constraint 4D-Var in ECMWF's Integrated Forecasting System (IFS), highlighting the potential of ML to systematically reduce model errors throughout the atmospheric column. Watt-Meyer et al. (2021) used random forests trained on FV3GFS nudging tendencies to correct model tendencies, achieving stable year-long runs and improved short-term forecasts for 500 hPa height, surface pressure, and near-surface temperature. Bretherton et al. (2022) corrected coarse-grid model errors by applying ML-learned temperature and humidity tendencies from a high-resolution reference, significantly improving prediction skills and precipitation patterns. Chen et al. (2022) used ML to learn the the analysis increments (i.e., the differences between the analysis and background, Evense and correct state-dependent model errors in NOAA's FV3-GFS. The online application of these corrections during model integration led to enhanced DA performance and improved 10-day predictions. Gregory et al. (2024) developed a hybrid dynamical–statistic framework that employs convolutional neural networks trained on sea ice concentration (SIC) assimilation increments, leading to improved five-year sea ice simulations. Most recently, Farchi et al. (2025) implemented an ML–based model error correction scheme within ECMWF's operational IFS. Their results indicated that offline-trained networks can already offer robust corrections, while online updates further enhance adaptability under diverse conditions. However, the potential benefits of ML-based model error correction for climate prediction ~~within an idealized framework. Our primary objective is to explore how the combination of~~ across different time scales remain largely unexplored. This is primarily due to the ~~data-driven error predictor and the dynamical model can enhance climate prediction as a function of lead time. Furthermore, in the coupled atmosphere-ocean model , the effects of errors in different components of the model in climate prediction are not yet fully understood. We aim to identify when correcting atmospheric errors or oceanic errors plays a pivotal role in improving climate prediction at different time scales~~sparsity of long-term observational records (such as those spanning the 20th century) in both time and space, which presents significant challenges for developing effective ML-based error correction models for climate prediction applications.

In this study, we investigate the potential of ML-based model error correction for climate prediction within an idealized framework. To this end, we adopt the hybrid modeling approach introduced by Brajard et al. (2021), which is based on MAOOAM. The ML-based error correction model aims to learn and correct dynamical climate model errors using analysis increments. Unlike Brajard et al. (2021), we conduct ensemble predictions with imperfect initial conditions (Farchi et al., 2023) , which better reflect realistic prediction scenarios (Wang et al., 2019; Bethke et al., 2021). Specifically, we examine how the effectiveness of ML-based error correction varies across different climate time scales. Moreover, given that the respective roles of atmospheric and oceanic errors in limiting climate predictability are not fully understood, we assess the relative contributions of these components to the overall prediction error.

The article is organized as follows: ~~Section~~ section 2 introduces the main methodological aspects of the study. Section 3 shows the prediction skill of the hybrid model compared with the dynamical climate model and discusses factors affecting the prediction skill of the hybrid model. Finally, a brief concluding summary is presented in section 4.

2 Methodology

In this study, we restrict our scope to model errors stemming solely from coarse resolutions in the atmospheric component. In this section, we describe the model (section 2.1), DA technique (section 2.2), and the ML approach (section 2.3). Rather than focusing on methodological developments, our goal is to examine how the advantages of ML-based error correction evolve in time in the context of climate prediction and to determine which errors should be corrected at different timescales. Further details in experiments are provided in section 2.4.

2.1 Modular Arbitrary-Order Ocean-Atmosphere Model

We utilize MAOOAM developed by De Cruz et al. (2016) in our study. MAOOAM consists of a two-layer ~~quasi-geostrophic (QG)~~ QG atmospheric component coupled with a QG shallow-water oceanic component. The coupling between these components incorporates wind forcings, ~~and~~ radiative and heat exchanges, enabling it to simulate climate variability. MAOOAM has been widely employed in qualitative analyses for various purposes (e.g., Penny et al., 2019; Brajard et al., 2021). Moreover, MAOOAM's numerical efficiency allows us ~~the execution of~~ to execute numerous climate prediction experiments at a relatively low computational cost.

In MAOOAM, the model variables are represented in terms of spectral modes. Specifically, d_{ax} (d_{ox}) represents the x-direction resolution, and d_{ay} (d_{oy}) represents the y-direction resolution in the atmosphere (ocean). The model state comprises n_a ($n_a = d_{ay}(2d_{ax}+1)$) modes of the atmospheric streamfunction ψ_a and temperature anomaly θ_a, as well as n_o ($n_o = d_{oy}d_{ox}$) modes of the oceanic streamfunction ψ_o and temperature anomaly θ_o. Consequently, the model state can be expressed as:

$$\mathbf{x} = (\psi_{a,1}, \psi_{a,2}, ..., \psi_{a,n_a}, \theta_{a,1}, \theta_{a,2}, ..., \theta_{a,n_a}, \psi_{o,1}, \psi_{o,2}, ..., \psi_{o,n_o}, \theta_{o,1}, \theta_{o,2}, ..., \theta_{o,n_o}) \tag{1}$$

The total number of variables in the model state is $2n_a + 2n_o$. ~~It is~~ Note that n_a is typically larger than n_o, reflecting the distinct characteristics of the two components in MAOOAM. The atmosphere exhibits faster dynamics and smaller-scale variability, necessitating a greater number of modes to adequately capture its behavior. In contrast, the ocean evolves more slowly and is dominated by larger-scale processes, which can be effectively represented using fewer modes (De Cruz et al., 2016). It is also important to note that variables with lower indices correspond to low-order (~~large scale~~large-scale) processes, while variables with higher indices correspond to high-order (~~small scale~~small-scale) processes. ~~One of the key features of MAOOAM is its ability to modify~~ Like many other models formulated in spectral space, MAOOAM offers flexibility in adjusting the number of atmospheric and oceanic ~~model variables simply by adjusting the model 's resolution in the x-direction or y-direction~~variables by simply modifying the model resolution in spectral space.

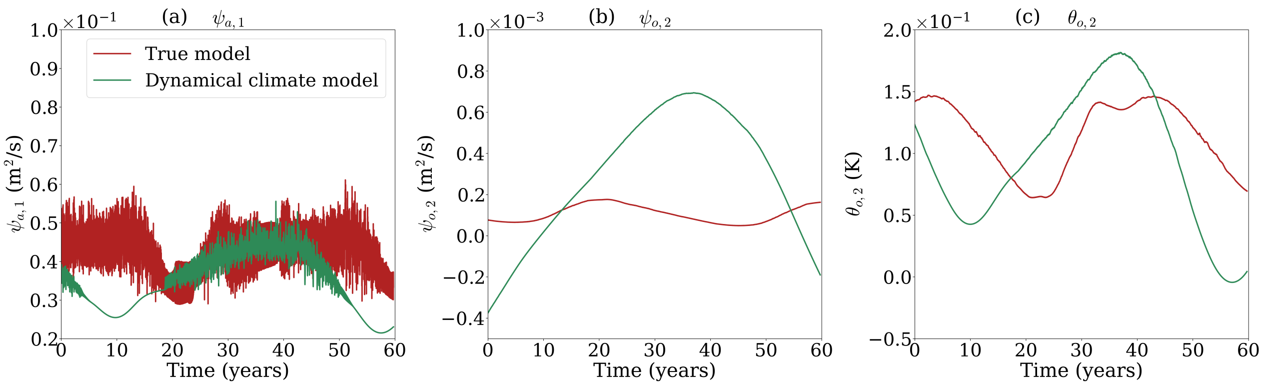

Figure 1. ~~The attractors in spectral space for (a)~~ Time series of the true model ~~: M56 and~~ (~~b~~red lines) and the dynamical climate model (green lines) for three key variables: ~~M36~~(a) $\psi_{a,1}$, (b) $\psi_{o,2}$, and (c) $\theta_{o,2}$.

In this study, we utilize two different configurations of MAOOAM: one denoted as **M56** and the other as **M36**. The M56 configuration comprises a total of 56 variables, with 20 atmospheric modes ($n_a = 20$) and 8 oceanic modes ($n_o = 8$). Specifically, the atmosphere in M56 operates at a 2x-4y (i.e., $d_{ax} = 2$ and $d_{ay} = 4$) resolution, and the ocean operates at a 2x-4y (i.e., $d_{ox} = 2$ and $d_{oy} = 4$) resolution. On the other hand, ~~The~~ the M36 configuration includes 36 variables, with 10 atmospheric modes ($n_a = 10$) and 8 oceanic modes ($n_o = 8$), identical to M56. The atmospheric component in M36 operates at a 2x-2y resolution ($d_{ax} = 2$, $d_{ay} = 2$), while the ocean component matches that of M56. Figure ~~?? displays the attractors of the~~ 1 displays time series of three key variables in ~~our~~ the true model M56 and ~~our dynamical~~ the dynamical climate model M36 in ~~the~~ spectral space, ~~showing they evolve differently~~ illustrating their different evolution patterns (De Cruz et al., 2016).

It is important to note that the key distinction between M36 and M56 lies in the atmosphere, where M36 has a reduced number of atmospheric modes, specifically 10 ~~mode less~~ modes fewer than M56 in the y-direction. This difference leads to a lack of higher-order atmospheric modes in M36, thereby unable to capture small-scale variability. The atmospheric error ~~could then propagate to all the components and variables of the system~~ in the y-direction propagates to the atmosphere in the x-direction and the ocean component through the coupling terms in the equations. Consequently, the primary source of model error in this study is attributed to the coarse resolution of ~~atmospheric part of the model~~the atmospheric component in the y-direction.

2.2 Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a flow-dependent and multivariate DA method and has been implemented for climate prediction ~~(e.g., Karspeck et al., 2013; Wang et al., 2019; Zhang et al., 2007)~~(e.g., Zhang et al., 2007; Karspeck et al., 2013; Wang et al., 2
. The EnKF constructs the background error covariance from the dynamical ensemble. The utilization of an ensemble-based error covariance ensures that the assimilation updates approximately respect to the model dynamics, thereby mitigating assimilation shocks (Evensen, 2003).

~~In this study , we utilize~~ All experiments in this study are conducted using the DAPPER package (Raanes, 2018)~~for conducting all experiments, as~~. The overall experimental setup is described in section 2.4 and depicted in Fig. 2. Specifically, we employ the finite-size ensemble Kalman filter (EnKF-N) method proposed by Bocquet et al. (2015). This method ~~reducing the amount of experimentation required in tuning the EnKF DA system, thereby~~ adaptively adjusts the inflation factor, thereby reducing the need for extensive manual tuning and enhancing the performance of the assimilation experiments~~, especially in case of the presence of model error, which we do in our setting~~. It is worth mentioning that we expect no significant alterations in the conclusions of this paper when using the traditional EnKF methods instead of EnKF-N.

2.3 Artificial Neural Network Architecture

We consider the dynamical climate model (described in section 2.1) in the following form:

$$\mathbf{x_{k+1}} = \mathcal{M}(\mathbf{x_k}), \tag{2}$$

where $\mathbf{x_{k+1}}$ represents the full model state at t_{k+1}, $\mathbf{x_k}$ represents the full model state at t_k and $\mathcal{M}$ represents the dynamical climate model integration from time t_k to t_{k+1}. The model error at time t_{k+1} is defined as:

$$\varepsilon_{k+1} = \mathbf{x^t_{k+1}} - \mathbf{x_{k+1}}, \tag{3}$$

where $\mathbf{x^t_{k+1}}$ represents the true state at time t_{k+1}.

We aim to use ANN to emulate the model error $\cancel{\varepsilon}\varepsilon_{k+1}$. Since the truth is not known in practice, the training of ANN ~~is using~~ uses the analysis increments produced by the EnKF ~~(Gregory et al., 2024).~~ (Brajard et al., 2021; Farchi et al., 2021b; Gregory et al., 2024). The architecture of ANN used in this study consists of four layers:

- The input layer includes a batch normalization layer (Ioffe, 2017), which helps to regularize and normalize the training process.
- The second layer is a dense layer with 100 neurons. It applies the rectified linear unit (ReLU) activation function, which introduces non-linearity into the network.
- The third layer has the same configuration as the second layer, with 50 neurons and ReLU activation function.
- The output layer, which is a dense layer with a linear activation function and produces the final predictions, is optimized using the “RMSprop” optimizer (Hinton et al., 2012) and includes an L2 regularization term with a value of 10^{-4}.

During training, the ANN model is trained with a batch size of 128 and for a total of 300 epochs.

The error surrogate model can be expressed as follows:

$$\varepsilon'_{k+1} = \mathcal{M}_{\mathrm{ANN}}(\mathbf{x_k}), \tag{4}$$

where $\mathcal{M}_{\mathrm{ANN}}$ represents the data-driven model built by ANN and ε'_{k+1} represents the model error estimated by ANN. The full state at time t_{k+1} of the hybrid model can be expressed as follows:

$$\mathbf{x^h_{k+1}} = \mathcal{M}(\mathbf{x_k}) + \mathcal{M}_{\mathrm{ANN}}(\mathbf{x_k}) \tag{5}$$

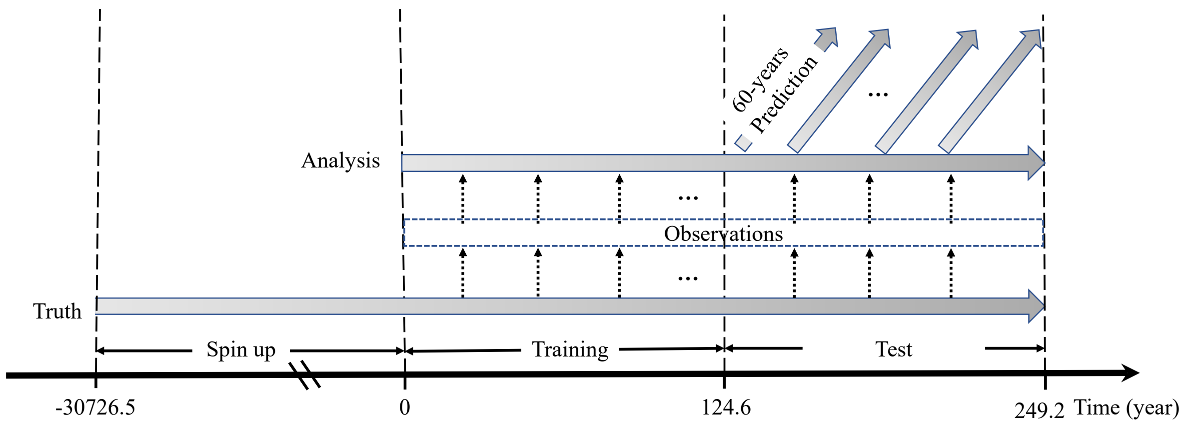

Figure 2. Schematic of experiments.

2.4 ~~Experimental~~ Experiment settings

We present the experimental setup in Fig. 2. The experiments are conducted using two configurations of MAOOAM, as described in section 2.1. The configuration with 56 variables (M56, section 2.1) represents the true climate system, while the configuration with 36 variables (M36, section 2.1) represents a dynamical climate prediction system. The experiments depicted in Fig. 2 are performed as follows:

- We integrate the M56 configuration with a time step of approximately 1.6 minutes for a spin-up period of 30726.5 years, as specified in De Cruz et al. (2016). Following the spin-up period, we continue the simulation for an additional 249 years, which we refer to as the “truth". To generate observations, we perturb the “truth" state using a Gaussian random noise. The standard deviation $(\boldsymbol{\sigma}^{\mathrm{hf}})$ of the noise is set to 10% of the temporal standard deviation of the true state ~~$\boldsymbol{\sigma}^{\mathrm{hf}}$~~ $(\mathbf{x}^{\mathbf{t}})$ after subtracting the one-month running average. Observations are generated ~~at intervals of approximately~~ every 27 hours in spectral space, while the observation operator H is the identity operator ($H = I$) and is also applied in spectral space.

- We assimilate synthetic observations into the dynamical climate model (M36) and generate ~~a reanalysis~~ analysis with 50 ensemble members over the same period ~~of~~ as the truth. The initial conditions of the ensemble are randomly sampled from a long free-run simulation of M36 after the spin-up period.

– We generate several sets of ensemble predictions with the dynamical climate model (M36) or the hybrid model. The prediction experiments start in each second year from the year 125 to the year 185, with each prediction lasting for 60 years. Each prediction consists of 50 ensemble members. The initial conditions for these ensembles are taken from the analysis (Fig. 2).

Note that both the observations and DA are conducted in the spectral space. Accordingly, the hybrid model is developed within the spectral space.

We split the analysis into two parts:

– Training data: The former 124.6 years of the dataset are used to train the ANN parameters to build the hybrid model (Fig. 2).

– Test data: The latter 124.6 years of the dataset are used to initialize prediction experiments (Fig. 2).

~~We utilized~~ It is worth noting that since we employ the same ANN configurations as ~~described in Brajard et al. (2021), although their study focused on a different objective within the MAOOAM framework. In our approach, the ANN parameters were trained in a single run of 300 epochs without incorporating validation data to adjust the ANN model during training . Upon completion of training, we analyzed~~ outlined in Brajard et al. (2021), the ANN parameters in this study are trained only once, without any modifications throughout the training process by using a separate validation set. We examined the loss curves ~~for both the training and test datasets. These loss curves confirmed~~ (not shown in this study) to assess the training behavior. The loss curves provided evidence that the network ~~continued to improve~~ was continuing to learn without signs of overfitting throughout the training process~~without signs of overfitting (not shown in the paper).~~ .

Brajard et al. (2021) focused on developing the hybrid model methodology; our study aims to explore the evolution of prediction skill as a function of lead time. We assess the prediction skill over a wider range of lead times, specifically up to 50 days for atmospheric variables and up to 60 years for oceanic variables. By examining the skill at various lead times, we can gain insights into the temporal evolution and long-term performance of the hybrid model, providing a more comprehensive understanding of its capabilities and limitations. To do so, our experimental setup is different from that of Brajard et al. (2021) in the following ways:

– We extended the simulation time to 219.2 years, while Brajard et al. (2021) generated an analysis dataset spanning 62 years for training, validation and testing. We divided our analysis dataset into two distinct parts: one for training the ANN and the other for testing purposes. This separation allows us to independently evaluate the performance of the trained ANN using data that was not used during the training phase.

– Our experiments utilize the analysis as initial conditions, while Brajard et al. (2021) uses perfect initial conditions (i.e., the truth) to initialize predictions. This choice reflects a more realistic scenario, as perfect knowledge of initial conditions is rarely available in the real framework. By using the analysis as initial conditions, we aim to capture the practical challenges associated with imperfect knowledge of the initial state in climate prediction.

– Our study incorporates an ensemble prediction strategy with 50 members, while Brajard et al. (2021) performed predictions using a single member (i.e., deterministic prediction). In the climate prediction community, probabilistic predictions based on ensembles are widely recognized. Ensembles provide a valuable means of quantifying uncertainty in climate predictions by generating multiple realizations rather than a single deterministic prediction.

2.5 Validation metrics

To evaluate the prediction skill of each variable, we employ the correlation and root mean square error (RMSE) skill score (RMSE-SS), which are commonly used metrics in weather forecasting and climate prediction. The correlation is defined as:

$$\text{Correlation} = \cancel{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}}\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}, \tag{6}$$

where x represents the prediction (ensemble mean) and y represents the truth. ~~n~~ N is the total number of prediction experiments and is equal to 30 (section 2.4).

The RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2}, \tag{7}$$

where x represents the prediction (ensemble mean), y represents the truth, and N is the total number of prediction experiments. The RMSE-SS compares the ~~root mean square error (RMSE)~~ RMSE of the prediction to the RMSE of a persistence prediction. It is defined as:

$$\cancel{\textbf{RMSE-SS}}\text{RMSE-SS} = 1 - \frac{\text{RMSE}_{\text{prediction}}}{\text{RMSE}_{\text{persistence}}}, \tag{8}$$

where $\text{RMSE}_{\text{prediction}}$ represents the RMSE between the prediction (ensemble mean) and the truth and $\text{RMSE}_{\text{persistence}}$ represents the RMSE between a persistence prediction (where the state remains the same as the initial conditions) and the truth. A positive RMSE-SS indicates that the prediction outperforms ~~the~~ persistence and demonstrates skill. On the other hand, a negative RMSE-SS indicates that the prediction performs worse than the persistence and lacks skill.

By utilizing the correlation and RMSE-SS, we can assess and compare the skill of the predictions generated by the dynamical climate model and the hybrid model across different variables within the same panel, as shown in Fig. 4~~(correlation) and Fig. ?? (RMSE-SS).~~.

To assess the statistical significance of the correlation and RMSE-SS~~results, we employ~~, we perform a two-tailed Student's t-test ~~. This statistical test helps determine if the prediction skill is statistically significant at different lead times.~~ based on the p-value. For the correlation, the null hypothesis is that the correlation is not significantly different from zero, implying no relationship between the predictions and truth. For RMSE-SS, we perform a hypothesis test to determine whether the squared errors (SE) from the prediction and persistence methods differ significantly. We compute the SE and use a two-tailed t-test to assess whether they are significantly different. Assuming sufficiently large sample sizes, the difference between the mean SEs can be approximated as normally distributed:

$$\mathrm{MSE}_{\mathrm{prediction}} - \mathrm{MSE}_{\mathrm{persistence}} \sim \mathcal{N}\left(0, \frac{s^2_{\mathrm{prediction}}}{N_{\mathrm{prediction}}} + \frac{s^2_{\mathrm{persistence}}}{N_{\mathrm{persistence}}}\right),$$

where $s^2_{\mathrm{prediction}}$ and $s^2_{\mathrm{persistence}}$ are the sample variances of the squared errors, and $N_{\mathrm{prediction}}$, $N_{\mathrm{persistence}}$ are the corresponding sample sizes. The resulting p-value represents the probability of observing the given difference (or larger) under the null hypothesis. A p-value below 0.05 is considered statistically significant, indicating that the prediction and persistence methods exhibit meaningfully different error characteristics.

To estimate the uncertainties of ~~the~~ correlation and RMSE-SS, we utilize the bootstrap method. We randomly select, with replacement, 30 data points from the 30 prediction experiments and calculate the correlation and RMSE-SS based on this sampled data. This procedure is repeated 10,000 times, resulting in a sample of 10,000 correlation and RMSE-SS values. The standard deviation of this sample is then used to estimate the uncertainties associated with the correlation and RMSE-SS. By conducting the t-test and utilizing the bootstrap method, we can obtain a more comprehensive understanding of the significance and reliability of the correlation and RMSE-SS values obtained from the prediction experiments.

In climate prediction, time-mean quantities such as monthly (Wang et al., 2019) or annual averages (Boer et al., 2016; Bethke et al., 2021) are often used because time averaging reduces the impact of chaotic weather variability, making the underlying climate signals more apparent. They also better meet the practical needs of sectors such as agriculture and energy, where planning is often based on mean conditions. In contrast, predicting higher-order statistics accurately remains challenging due to model limitations and computational costs, particularly when high-resolution Earth system models are required. Nevertheless, there is growing interest in representing complex statistical properties to improve the prediction of extreme events and support climate risk assessments.

To evaluate prediction skill across different time scales, we apply two complementary strategies. For short-term predictions, instantaneous outputs sampled every 27 hours are used to represent daily variations. For long-term predictions, model outputs are averaged annually to assess the ability to capture low-frequency variability.

3 Results

3.1 Prediction skill

The distinction between short-term (daily) and long-term (yearly-averaged) prediction scales in this study is based on the fundamentally different error growth characteristics of atmospheric and oceanic variables. As illustrated in Fig. 3, atmospheric variables exhibit rapid error amplification, with a doubling error time of approximately one day and saturation occurring within about ten days. In contrast, oceanic variables demonstrate much slower error growth, with errors roughly doubling over the first year and continuing to grow gradually over the subsequent decade.

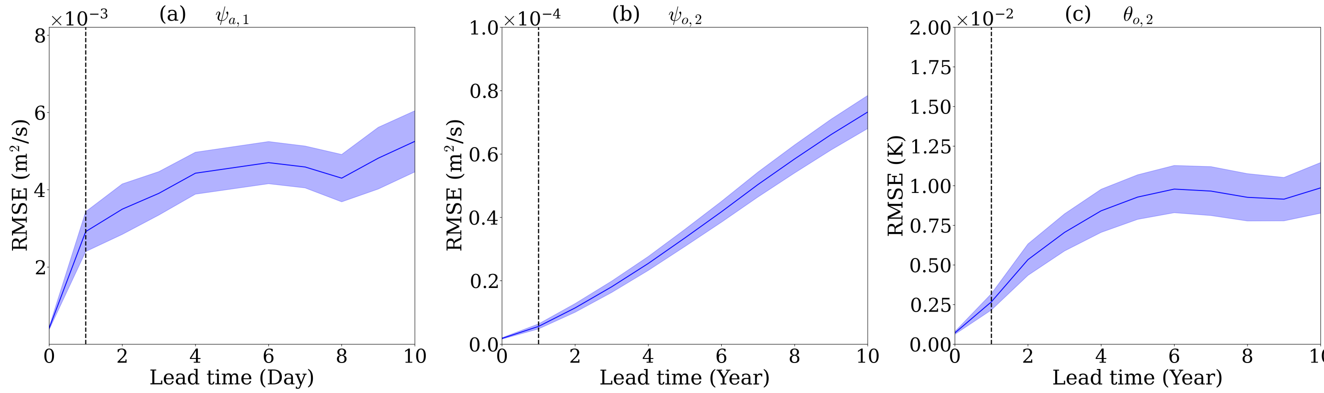

Figure 3. RMSE of dynamical prediction as a function of lead time for three key variables in spectral space: (a) $\psi_{a,1}$, (b) $\psi_{o,2}$, and (c) $\theta_{o,2}$. The atmospheric variable ($\psi_{a,1}$) is evaluated based on instantaneous outputs sampled every 27 hours, while the oceanic variables ($\psi_{o,2}$ and $\theta_{o,2}$) are evaluated using yearly averages. Shading indicates one standard deviation, representing the uncertainty of prediction skill, estimated using the bootstrap method. The vertical dashed lines represent the time of doubling error.

Within the coupled model framework, the hybrid model is developed to enhance prediction skill across both short-term and long-term timescales. To evaluate its performance, we adopt 50-day and 60-year prediction horizons as representative benchmarks for the subseasonal-to-seasonal and decadal prediction regimes, respectively. The 50-day prediction reflects the model's capability in capturing fast-evolving atmospheric processes, while the 60-year prediction assesses its capacity to maintain predictability over longer oceanic timescales.

~~Same as Fig. 4, but for RMSE-SS.~~

Figures 4a and ~~??a~~ 4c show respectively the correlation and RMSE-SS of the dynamical climate model for both atmospheric temperature θ_a and streamfunction ψ_a in the spectral space. We find that the variables in low-order atmospheric modes, such as $\psi_{a,2}$, $\psi_{a,3}$, $\theta_{a,2}$ and $\theta_{a,3}$, have significant prediction skills over 10 days. While most variables in high-order modes have significant skills within a few days, some do not have prediction skills all the time (i.e. $\psi_{a,9}$, $\psi_{a,10}$ and $\theta_{a,10}$). Figures 4b and ~~??b~~ 4d show the correlation and RMSE-SS of the hybrid model for atmospheric variables. For atmospheric temperature, the hybrid model is skillful for up to 50 days for most modes (Fig. 4b), with a significant reduction in prediction error beyond ~~ten~~ 10 days for most modes (Fig. ~~??b~~4d). For atmospheric streamfunction, the hybrid model is skillful in predicting low-order atmospheric modes for up to 50 days and high-order modes for up to 15 days. Overall, the hybrid model has higher correlations and RMSE-SS than the dynamical climate model for atmospheric variables. And the hybrid model exhibits greater improvements in lower-order modes compared to higher-order modes (~~Fig~~Figs. 4a~~and~~ , 4b, ~~Fig. ??a and ??b~~4c, and 4d).

~~In the coupled model, the purpose of introducing ML to correct model errors is not only to improve the short-term atmospheric prediction skills (e.g., less than 20 days) but also to improve the long-term prediction skills (e.g., over 10 years).~~

~~Figures 4c and ??c~~ Figures 4e and 4g show the correlation and RMSE-SS of the dynamical climate model for oceanic temperature and streamfunction. Since the ocean ~~has~~ exhibits slower variability than the atmosphere, ~~the dynamical model~~

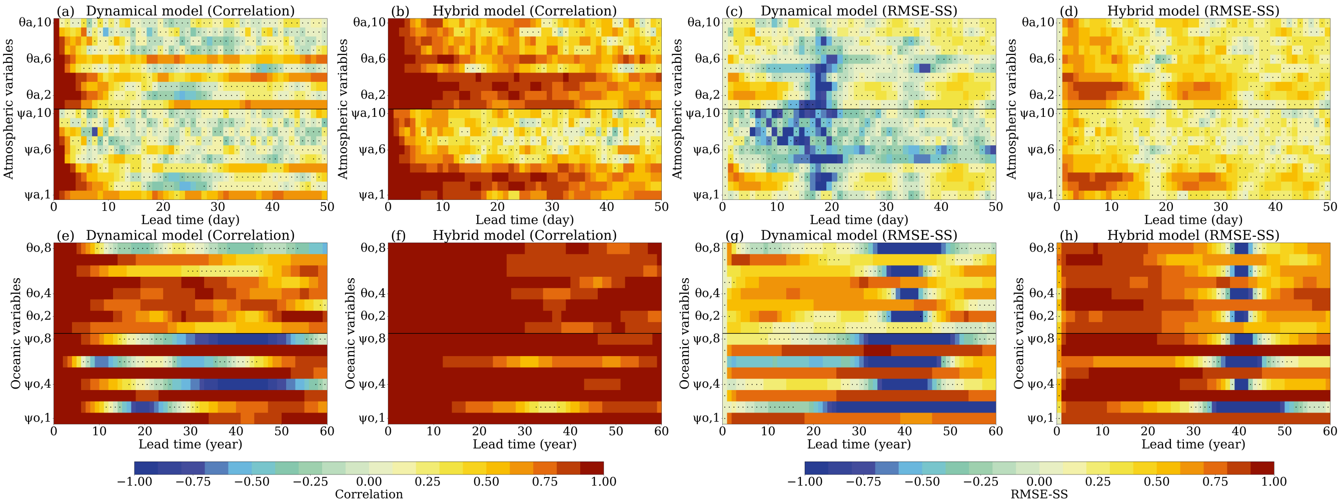

Figure 4. Correlation and RMSE-SS as a function of the prediction lead time for different variables. (a,~~c~~e) The correlation between the dynamical climate model and truth. (c,g) The RMSE-SS between the dynamical climate model and truth. (b,~~d~~f) The correlation between the hybrid model and truth. (d,h) The RMSE-SS between the hybrid model and truth. The atmospheric variables are calculated based on daily data, while the oceanic variables are based on annual average data. The black dot indicates the correlation does not exceed the 95% significance test.

~~has significant prediction skills~~ we compute annual means for oceanic variables to evaluate the model's prediction skill on interannual timescales. The dynamical climate model demonstrates significant prediction skill for up to 60 years in oceanic temperature ~~in most modesand~~ across most modes, and in oceanic streamfunction in ~~some~~ certain modes. Overall, odd-numbered modes exhibit higher ~~predictive~~ prediction skill than even-numbered modes, related to our experimental design (i.e., the difference in atmospheric y-direction mode resolution between M56 and M36). In addition, the oceanic temperature is more predictable than the oceanic streamfunction in the spectral space. Figures 4~~d and ??d~~ f and 4h present the prediction skills of the hybrid model. The hybrid model has significant prediction skills in both oceanic temperature and streamfunction in all modes for up to 60 years. It is worth noting that the hybrid model has higher correlations and RMSE-SS than the dynamical climate model, in particular, for oceanic temperature in the first and last modes and oceanic streamfunctions in some modes in which the dynamical climate model has no prediction skill at all (e.g., ~~$\varphi_{o,2}$ and $\varphi_{o,6}$~~ $\psi_{o,2}$ and $\psi_{o,6}$).

To further demonstrate the advantages of the hybrid model, we use a ~~ten-day~~ 10-day lead time for atmospheric variables and a ~~forty-year~~ 40-year lead time for oceanic variables as examples to show the prediction skills of the hybrid model in the physical space (Fig. 5 and Fig. 6~~for correlation, Fig. ?? and Fig. ?? for RMSE-SS~~).

~~Same as Fig. 5, but for RMSE-SS.~~

For atmospheric variables, both atmospheric streamfunction and temperature exhibit similar spatial characteristics (~~Figs. 5and ??~~Fig. 5). We find that the hybrid model has similar spatial patterns but outperforms the dynamical climate model in most grid points.

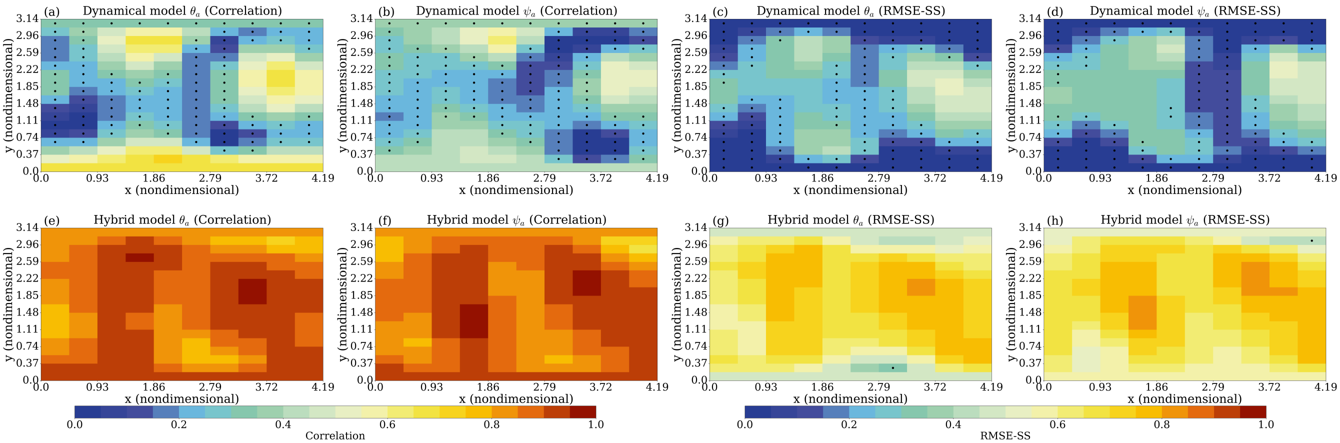

Figure 5. ~~Correlation~~ Spatial distributions of correlation and RMSE-SS at ~~the~~ prediction lead day 10 for atmospheric variables. Panels (a,~~b~~–d) ~~The correlation between~~ show results from the dynamical climate model: (a) correlation between predicted and ~~the truth for~~ observed atmospheric temperature~~,~~; (~~c,d~~b) ~~The~~ correlation ~~between the dynamical model~~ for atmospheric streamfunction; (c) RMSE-SS for atmospheric temperature; and ~~the truth~~ (d) RMSE-SS for atmospheric streamfunction. Panels (e–h) show corresponding results from the hybrid model. The black dot indicates the correlation and RMSE-SS does not exceed the 95% significance test.

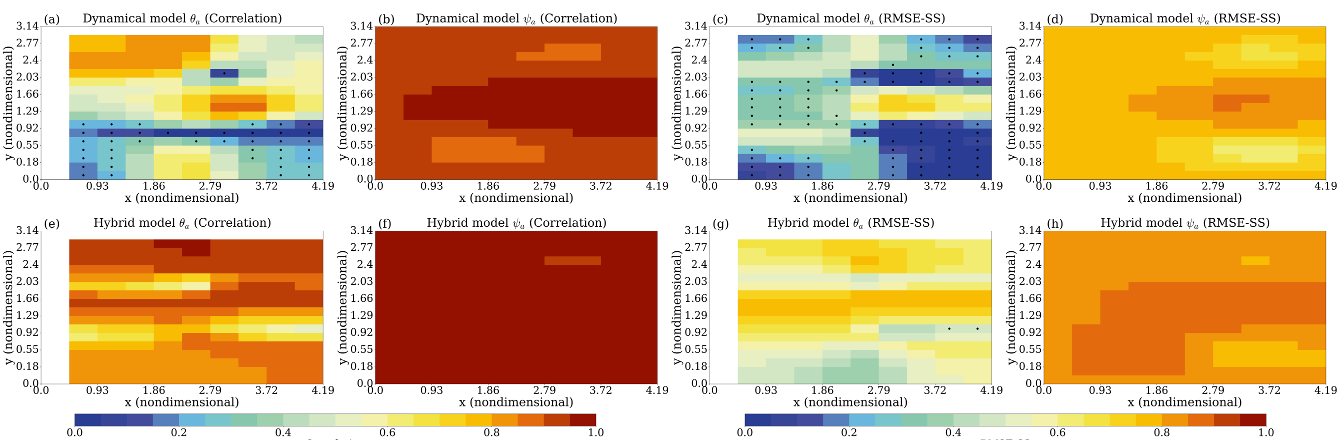

Figure 6. ~~Correlation~~ Spatial distributions of correlation and RMSE-SS at ~~the~~ prediction lead year 40 for oceanic variables. Panels (a,~~b~~–d) ~~The correlation between~~ show results from the dynamical climate model: (a) correlation between predicted and ~~the truth~~ observed oceanic temperature; (b) correlation for oceanic streamfunction~~,~~; (c~~,d~~) ~~the correlation between the dynamical model and the truth~~ RMSE-SS for oceanic temperature; and (d) RMSE-SS for oceanic streamfunction. ~~The white areas in the temperature map result~~ Panels (e–h) show corresponding results from ~~consistently zero temperature anomalies at~~ the ~~western and northern boundaries, which prevents the calculation of correlation~~hybrid model. The black dot indicates the correlation and RMSE-SS does not exceed the 95% significance test.

~~Same as Fig. 6, but for RMSE-SS.~~

~~Due to the slow nature of oceanic variability in the MAOOAM, at a lead time of 40 years, the dynamical model still maintains high predictive skills at all~~ For oceanic temperature, the dynamical climate model loses prediction skill over the majority of grid points (~~Figures~~ Fig. 6a and ~~??a). Compared to the dynamical model~~Fig. 6c). In contrast, the hybrid model ~~exhibits higher correlations and RMSE-SS at all grid points , outperforming the dynamical model (Figures 6b and ??d~~demonstrates significantly higher prediction skill across most grid points (Figs. 6e and 6g). For oceanic ~~temperature, the dynamical model has higher correlations and significant RMSE-SS in the northwest region~~streamfunction, owing to the slow nature of variability in MAOOAM, the dynamical climate model retains high prediction skill at all grid points even at a 40-year lead time (Figs 6b and 6d). The hybrid model ~~shows higher prediction skills which are statistically significant at most grids. For example, the dynamical model lacks prediction skills in the northeast region, while the hybrid model has high skills (Figures 6d and ??d~~further improves upon this, showing higher correlations and RMSE-SS at all grid points, thereby outperforming the dynamical climate model (Figs. 6f and 6h).

For long-term climate prediction, there are additional requirements that the hybrid model must meet. Specifically, the model should be capable of running for extended periods without diverging or exhibiting significant physical instability. In our study, we find that the hybrid model maintains stability and does not experience significant physical instability during the 60-year prediction period.

In summary, the overall performance of the hybrid model surpasses that of the dynamical climate model in both spectral and physical space, demonstrating the advantages of incorporating a data-driven error correction model constructed by ~~the~~ ML. This result highlights the potential benefits of leveraging data-driven approaches to improve ~~dynamical~~ climate prediction skills.

3.2 Importance of atmospheric or oceanic error correction~~in climate prediction~~

In this section, we extend our analysis by constructing two additional hybrid models to explore the influence of correcting atmospheric and oceanic errors separately. These models are trained using the same inputs as in the previous section, but are designed to correct either atmospheric errors or oceanic errors. By comparing the prediction skills of the regional averaged variables in physical space among these hybrid models, we ~~aim to determine which error is more important for prediction on different time scales. Through this analysis, we gain insights~~ gain some insight into the relative importance of atmospheric and oceanic error correction for the overall ~~prediction performance~~ performance of the climate prediction on different time scales.

~~Same as Fig. 7, but for RMSE-SS.~~

~~In Figures~~ In Figs. 7a, 7b, ~~??a, and ??b~~7e, and 7f, we present the correlation and RMSE-SS of different models specifically for the atmospheric streamfunction and temperature. We observe that there is minimal difference in prediction skill between correcting only the atmospheric errors (green line) and correcting both the atmospheric and oceanic errors (red line). However, in the early ~~forecast~~ prediction period (less than 20 days), correcting only atmospheric errors has slightly higher skills than correcting both atmospheric and oceanic errors simultaneously. When comparing the hybrid models with the dynamical climate model (blue line), we find that correcting only the oceanic errors (cyan line) does not lead to improvements in atmospheric

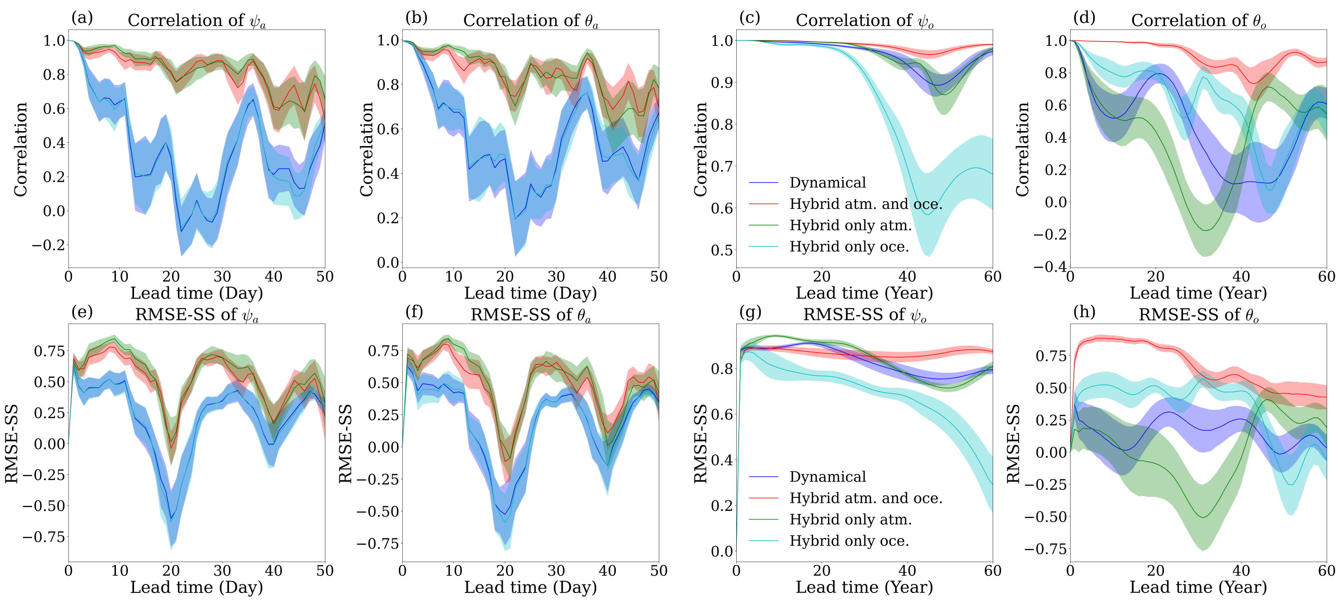

Figure 7. Correlation (a-d) and RMSE-SS (e-h) as a function of lead time (50 days for the atmospheric variable and 60 years for the oceanic variables). Shading shows one standard deviation calculated by the bootstrap method described in section 2.5. The red line is the correlation/RMSE-SS of the hybrid model built by correcting both atmospheric and oceanic model errors, the green line is the correlation/RMSE-SS of the hybrid model built by only correcting atmospheric model errors, the cyan line is the correlation/RMSE-SS of the hybrid model built by only correcting oceanic model errors and the blue line is the correlation/RMSE-SS of the dynamical climate model.

prediction. It is related to the fact in MAOOAM that the atmosphere mostly drives the ocean but the ocean has too weak ~~influences in~~ influence on the atmosphere for short-term climate prediction (Jung and Vitart, 2006).

In ~~Figures~~ Figs. 7c, 7d, ~~??c and ??d~~7g and 7h, we focus on the long-term prediction skill of various hybrid models for the oceanic streamfunction and temperature. Our results reveal that the highest prediction skill over 60 years is achieved when both atmospheric and oceanic errors are corrected (red line). The hybrid models constructed by correcting only atmospheric or oceanic model errors exhibit different performances. For the oceanic streamfunction (Fig. 7c~~), solely correcting~~ ,g), correcting only oceanic errors (cyan line) does not improve ~~the~~ prediction skill. ~~Specifically, as the~~ As lead time increases, ~~it exhibits lower skills compared to the dynamical~~ both the correlation and RMSE-SS metrics indicate a degradation in performance, with skill levels even lower than the dynamical climate model (blue line). ~~When only correcting~~ In contrast, correcting only atmospheric errors (green line) ~~, the improvement in prediction skill occurs in~~ significantly improves prediction skill within the first 20 years~~of lead time~~. However, ~~after~~ beyond 20~~years, while the skill of correcting atmospheric errors starts to decline~~ –30 years, the skill gradually declines and becomes comparable to ~~the skill~~ that of the dynamical ~~model (blue line), simultaneously correcting~~ climate model. Notably, the hybrid correction that simultaneously addresses both atmospheric and oceanic model errors ~~is still better than the dynamical model. For RMSE-SS (Fig. ??e), correcting only oceanic errors (cyan line) also consistently yields the lowest RMSE-SS compared to other predictions, including the dynamical model (blue line). Correcting only atmospheric~~

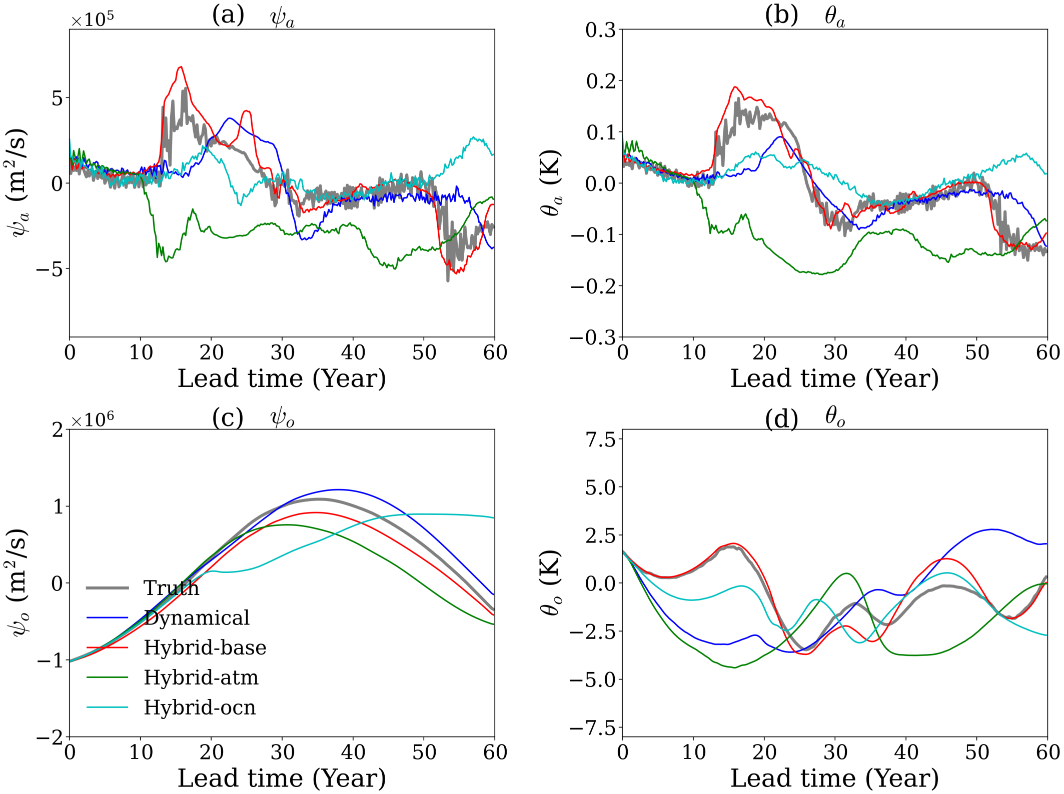

Figure 8. A case study based on the ensemble mean and monthly mean average illustrating the simulation results of four variables averaged over the whole domain in the physical space. (a) atmospheric streamfunction, (b) atmospheric temperature, (c) oceanic streamfunction, (d) oceanic temperature.

~~errors (green line) achieves the best RMSE-SS within the first 20 years of lead time. However, while the skill of correcting atmospheric errors alone begins to decline~~ (red line) consistently outperforms the dynamical climate model after 30 years, ~~reaching a level similar to that of the dynamical model, the hybrid model that simultaneously corrects both atmospheric and oceanic errors (red line) continues to outperform the dynamical model~~in both correlation and RMSE-SS metrics.

Regarding oceanic temperature (Figs. 7d and ~~??d~~7h), correcting only atmospheric errors does not improve the prediction of oceanic temperatures, while only correcting oceanic errors can enhance the prediction skill of oceanic temperatures. Additionally, simultaneously correcting both atmospheric and oceanic errors (red line) can achieve the highest prediction skills ~~all the lead time~~at all lead times.

To better illustrate the advantages of the hybrid model, we use ~~a set of experimental results~~ one prediction experiment as an example to demonstrate the benefits of correcting model errors for long-term simulations (Fig. 8). For atmospheric variables (Fig. 8a and 8b), correcting only one component does not effectively simulate the slow frequency atmospheric processes (i.e., low-frequency signals around lead time 20 years), while simultaneously correcting both atmospheric and oceanic model errors (red lines) can better capture this variation. For the oceanic streamfunction (Fig. 8c), solely correcting oceanic errors (cyan lines) causes a phase change compared to the truth. However, the phase of the other models still matches the truth, with some differences in magnitude and timing. For oceanic temperature (Fig. 8d), correcting only atmospheric errors leads to the largest deviation from the truth (grey lines) in the first 20 years, which is similar to the dynamical climate model. Correcting the oceanic errors is better, but still poorer than correcting both atmospheric and oceanic errors (red lines), which leads to predictions very close to the truth.

In summary, for short-term atmospheric predictions, correcting atmospheric model errors yields better results, while for long-term simulations, correcting both oceanic and atmospheric errors provides the best predictions.

4 Summary and discussions

In this study, we applied a method to online correct the error in a simplified atmosphere-ocean coupled model (MAOOAM). The errors in the MAOOAM setup stem from resolution limitations in the atmospheric ~~model~~component. We constructed a data-driven predictor of dynamical climate model error with ~~the~~ ML techniques and integrated it with the dynamical climate model, creating a hybrid statistical-dynamical model. By incorporating the model error correction through the hybrid model, we significantly enhanced the prediction skills for both atmospheric and oceanic variables at different lead times in both spectral and physical space. This approach allowed us to mitigate the limitations of the dynamical climate model and achieve more accurate climate predictions.

~~We also investigated the impact of individually correcting either atmospheric or~~ This study also examined the respective impacts of correcting atmospheric and oceanic model errors on prediction ~~skills. For~~ skill. Our results indicate that short-term atmospheric ~~prediction, its accuracy is more~~ predictions are primarily influenced by atmospheric model errors, while ~~only~~ correcting oceanic errors ~~has little impact on short-term atmospheric prediction Balmaseda and Anderson (2009)~~alone has a limited effect (e.g., Balmaseda and Anderson, 2009). For long-term ocean prediction, correcting ~~the atmospheric model error is critical for oceanic streamfunction prediction, since it is sensitive to atmospheric forcings. For oceanic temperature, correcting the oceanic model error is more important due to the long memory of ocean heat content Griffies et al. (2015). Our findings suggested correcting atmospheric errors for short-term atmosphere prediction while correcting both atmospheric and~~ atmospheric errors is essential due to their role in surface forcing, while correcting oceanic ~~model errors for long-term climate prediction~~ errors plays a more critical role in predicting ocean temperature. It is worth noting that in our experiment setup, the ocean component is perfect, and its prediction errors primarily come from the errors in the atmospheric component. However, correcting the ocean model errors can influence the atmosphere through the coupling between the ocean and the atmosphere. Although the experimental setup is not ideal, our results still provide some insights into the relative importance of oceanic error correction for the prediction on different time scales.

This study serves as a proof of concept, demonstrating the potential of using ML to learn and correct errors in dynamical climate models, thereby enhancing their prediction skills. Although conducted in the simplified atmosphere-ocean coupled model MAOOAM, this study contributes to the understanding of the impact of correcting model errors on climate prediction in the atmosphere-ocean coupling process. It emphasizes the importance of errors in different components of coupled models and highlights how correcting errors in various components can improve predictions on different time scales. Future applications involve applying this method to realistic climate models, which are inherently more complex than MAOOAM, and exploring the prediction skills under such conditions.

Code and data availability. All data used in this study are generated by the experiments in section 2.4 and are available at https://doi.org/10.5281/zenodo.7725687. And the code is available at https://github.com/zikanghe/MAOOAM-hybrid-papaer.

Author contributions. Conceptualization: ZH, YW, JB. Analysis and Visualization: ZH. Interpretation of results: ZH, YW, JB. Writing (original draft): ZH, YW. Writing (reviewing and editing original draft): ZH, YW, JB, XW, ZS.

Competing interests. The author declares that no competing interests.

Acknowledgements. This study was funded by the National Key $R\&D$ Program of China (2022YFE0106400), ~~the China Scholarship Council (202206710071),~~ Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_0657,KYCX24_0808), the China Scholarship Council (202206710071), the Special Founds for Creative Research (2022C61540), the Opening Project of the Key Laboratory of Marine Environmental Information Technology (521037412). YW was funded by the Research Council of Norway (Grant nos. 328886, 309708) and the Trond Mohn Foundation under project number BFS2018TMT01. JB was funded by the Reseach Council of Norway (Grant no. 309562). ZS was funded by the National Natural Science Foundation of China (42176003), the Fundamental Research Funds for the Central Universities (B210201022). This work received grants for computer time from the Norwegian Program for supercomputer (NN9039K) and storage grants (NS9039K).

References

Balmaseda, M. and Anderson, D.: Impact of initialization strategies and observations on seasonal forecast skill, Geophysical research letters, 36, 2009.

Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., Samuelsen, A., Langehaug, H., Svendsen, L., Chiu, P.-G., et al.: NorCPM1 and its contribution to CMIP6 DCPP, Geoscientific Model Development, 14, 7073–7116, 2021.

Bocquet, M., Raanes, P. N., and Hannart, A.: Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation, Nonlinear Processes in Geophysics, 22, 645–662, 2015.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., et al.: The decadal climate prediction project (DCPP) contribution to CMIP6, Geoscientific Model Development, 9, 3751–3777, 2016.

Bonavita, M. and Laloyaux, P.: Machine learning for model error inference and correction, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 232, 2020.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization, Philosophical Transactions of the Royal Society A, 379, 20200 086, 2021.

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., Perkins, W. A., Clark, S. K., and Harris, L.: Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 794, 2022.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change, 9, e535, 2018.

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., and Tulich, S.: Correcting Systematic and State-Dependent Errors in the NOAA FV3-GFS Using Neural Networks, Journal of Advances in Modeling Earth Systems, 14, 2022.

De Cruz, L., Demaeyer, J., and Vannitsem, S.: The modular arbitrary-order ocean-atmosphere model: MAOOAM v1. 0, Geoscientific Model Development, 9, 2793–2808, 2016.

Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L., and Van Oldenborgh, G.: Initialized near-term regional climate change prediction, Nature communications, 4, 1715, 2013a.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.: Seasonal climate predictability and forecasting: status and prospects, Wiley Interdisciplinary Reviews: Climate Change, 4, 245–268, 2013b.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.

Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., and Malartic, Q.: A comparison of combined data assimilation and machine learning methods for offline and online model error correction, Journal of computational science, 55, 101 468, 2021a.

Farchi, A., Laloyaux, P., Bonavita, M., and Bocquet, M.: Using machine learning to correct model error in data assimilation and forecast applications, Quarterly Journal of the Royal Meteorological Society, 147, 3067–3084, 2021b.

Farchi, A., Chrust, M., Bocquet, M., Laloyaux, P., and Bonavita, M.: Online model error correction with neural networks in the incremental 4D-Var framework, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 474, 2023.

Farchi, A., Chrust, M., Bocquet, M., and Bonavita, M.: Development of an offline and online hybrid model for the Integrated Forecasting System, Quarterly Journal of the Royal Meteorological Society, p. e4934, 2025.

Gregory, W., Bushuk, M., Zhang, Y., Adcroft, A., and Zanna, L.: Machine learning for online sea ice bias correction within global ice-ocean simulations, Geophysical Research Letters, 51, e2023GL106 776, 2024.

Griffies, S. M., Winton, M., Anderson, W. G., Benson, R., Delworth, T. L., Dufour, C. O., Dunne, J. P., Goddard, P., Morrison, A. K., Rosati, A., et al.: Impacts on ocean heat from transient mesoscale eddies in a hierarchy of climate models, Journal of Climate, 28, 952–977, 2015.

Hinton, G., Srivastava, N., and Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, Cited on, 14, 2, 2012.

Ioffe, S.: Batch renormalization: Towards reducing minibatch dependence in batch-normalized models, Advances in neural information processing systems, 30, 2017.

Jung, T. and Vitart, F.: Short-range and medium-range weather forecasting in the extratropics during wintertime with and without an interactive ocean, Monthly weather review, 134, 1972–1986, 2006.

Karspeck, A. R., Yeager, S., Danabasoglu, G., Hoar, T., Collins, N., Raeder, K., Anderson, J., and Tribbia, J.: An ensemble adjustment Kalman filter for the CCSM4 ocean component, Journal of Climate, 26, 7392–7413, 2013.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., Danabasoglu, G., Dirmeyer, P. A., Doblas-Reyes, F. J., Domeisen, D. I., et al.: Current and emerging developments in subseasonal to decadal prediction, Bulletin of the American Meteorological Society, 101, E869–E896, 2020.

Moufouma-Okia, W. and Jones, R.: Resolution dependence in simulating the African hydroclimate with the HadGEM3-RA regional climate model, Climate Dynamics, 44, 609–632, 2015.

Palmer, T. and Stevens, B.: The scientific challenge of understanding and estimating climate change, Proceedings of the National Academy of Sciences, 116, 24 390–24 395, 2019.

Palmer, T. N.: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models, Quarterly Journal of the Royal Meteorological Society, 127, 279–304, 2001.

Penny, S., Bach, E., Bhargava, K., Chang, C.-C., Da, C., Sun, L., and Yoshida, T.: Strongly coupled data assimilation in multiscale media: Experiments using a quasi-geostrophic coupled model, Journal of Advances in Modeling Earth Systems, 11, 1803–1829, 2019.

Penny, S. G. and Hamill, T. M.: Coupled data assimilation for integrated earth system analysis and prediction, Bulletin of the American Meteorological Society, 98, ES169–ES172, 2017.

Raanes, P. N.: nansencenter/DAPPER: Version 0.8, https://doi.org/10.5281/zenodo.2029296, 2018.

Richter, I.: Climate model biases in the eastern tropical oceans: Causes, impacts and ways forward, Wiley Interdisciplinary Reviews: Climate Change, 6, 345–358, 2015.

Richter, I. and Tokinaga, H.: An overview of the performance of CMIP6 models in the tropical Atlantic: mean state, variability, and remote impacts, Climate Dynamics, 55, 2579–2601, 2020.

Tian, B. and Dong, X.: The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation, Geophysical Research Letters, 47, e2020GL087 232, 2020.

Wang, Y., Counillon, F., Keenlyside, N., Svendsen, L., Gleixner, S., Kimmritz, M., Dai, P., and Gao, Y.: Seasonal predictions initialised by assimilating sea surface temperature observations with the EnKF, Climate Dynamics, 53, 5777–5797, 2019.

Watson, P. A.: Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction, Journal of Advances in Modeling Earth Systems, 11, 1402–1417, 2019.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., and Bretherton, C. S.: Correcting weather and climate models by machine learning nudged historical simulations, Geophysical Research Letters, 48, e2021GL092 555, 2021.

Williamson, D. L., Drake, J. B., Hack, J. J., Jakob, R., and Swarztrauber, P. N.: A standard test set for numerical approximations to the shallow water equations in spherical geometry, Journal of computational physics, 102, 211–224, 1992.

Zhang, S., Harrison, M., Rosati, A., and Wittenberg, A.: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies, Monthly Weather Review, 135, 3541–3564, 2007.