https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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. Despite notable advances in dynamical climate modeling that have improved our understanding of climate variability, these systematic errors can still degrade predictive 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 model error, which is then integrated with the dynamical model to form a hybrid system. Comparison experiments show that the hybrid model consistently outperforms the standalone dynamical 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, 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.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

Dynamical 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 models and model errors are two critical sources that limit the prediction skill of dynamical 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 models to estimate best the state of the climate system (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 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 model. ML can be employed to construct a data-driven predictor of model errors, which can then be integrated with the dynamical model to create a hybrid statistical-dynamical model (e.g., Watson, 2019; Farchi et al., 2021; 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., 2021) have 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. (2021) 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 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 conditions are 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, 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.

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 of ML-based model error correction for climate prediction within an idealized framework. Our primary objective is to explore how the combination of the data-driven error predictor and the dynamical model can enhance climate prediction as

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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.

The article is organized as follows: Section 2 introduces the main methodological aspects of the study. Section 3 shows the prediction skill of the hybrid model compared with the dynamical 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) 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 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:

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

The total number of variables in the model state is $2n_a + 2n_o$. It is important to note that variables with lower indices correspond to low-order (large scale) processes, while variables with higher indices correspond to high-order (small scale) processes. One of the key features of MAOOAM is its ability to modify the number of atmospheric and oceanic model variables simply by adjusting the model's resolution in the x-direction or y-direction.

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.,

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

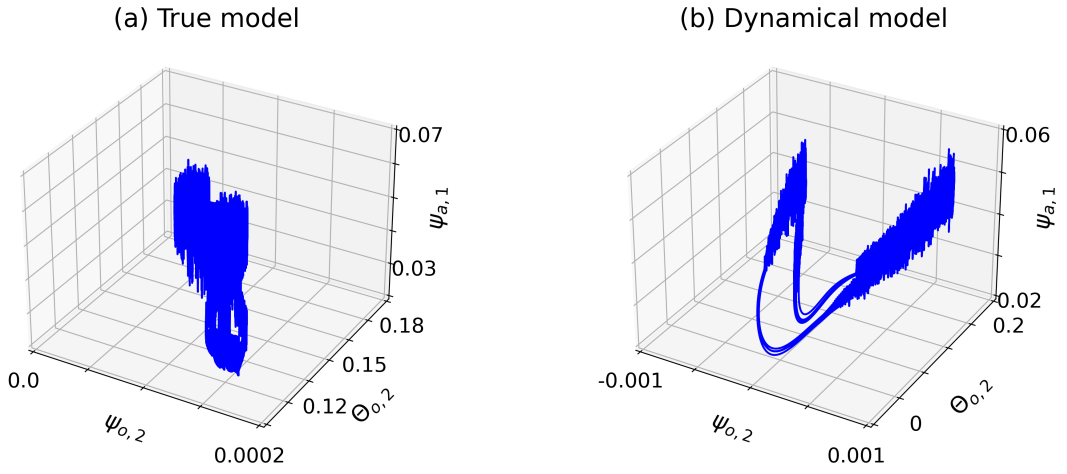

Figure 1. The attractors in spectral space for (a) the true model: M56 and (b) the dynamical model: M36.

$d_{ox} = 2$ and $d_{oy} = 4$) resolution. On the other hand, 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 1 displays the attractors of the three key variables in our true model M56 and our dynamical model M36 in the spectral space, showing they evolve differently (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 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 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.

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). 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 the DAPPER package (Raanes, 2018) for conducting all experiments, as 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 enhancing the performance of the assimilation experiments, especially in case of the presence of model error, which we do in our setting. It

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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 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 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 ε. Since the truth is not known in practice, the training of ANN is using the analysis increments produced by the EnKF (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:

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

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

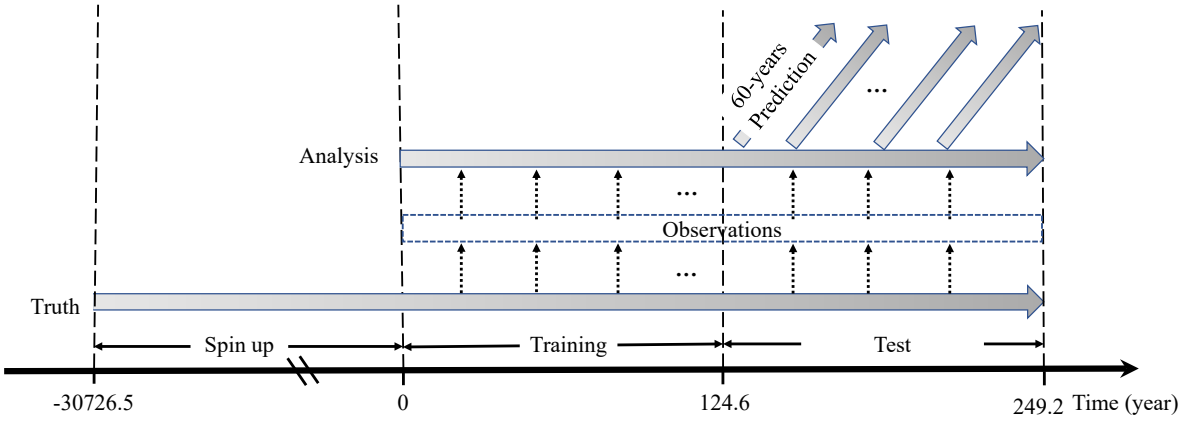

Figure 2. Schematic of experiments.

2.4 Experimental 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 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 of the noise is set to 10% of the temporal standard deviation of the true state $\boldsymbol{\sigma}^{\mathrm{hf}}$ after subtracting the one-month running average. Observations are generated at intervals of approximately 27 hours.

- We assimilate synthetic observations into the dynamical model (M36) and generate a reanalysis with 50 ensemble members over the same period of 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 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

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

prediction consists of 50 ensemble members. The initial conditions for these ensembles are taken from the analysis (Fig. 2).

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 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 the loss curves for both the training and test datasets. These loss curves confirmed that the network continued to improve throughout the training process without signs of overfitting (not shown in the paper).

2.5 Validation metrics

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

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

where x represents the prediction (ensemble mean) and y represents the truth. n is the total number of prediction experiments.

The RMSE-SS compares the root mean square error (RMSE) of the prediction to the RMSE of a persistence prediction. It is defined as:

$$\textbf{RMSE-SS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{prediction}}}{\mathrm{RMSE}_{\mathrm{persistence}}}, \tag{7}$$

where $\mathrm{RMSE}_{\mathrm{prediction}}$ represents the RMSE between the prediction (ensemble mean) and the truth and $\mathrm{RMSE}_{\mathrm{persistence}}$ 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 model and the hybrid model across different variables within the same panel, as shown in Fig. 3 (correlation) and Fig. 4 (RMSE-SS).

To assess the significance of the correlation and RMSE-SS results, we employ a two-tailed Student's t-test. This statistical test helps determine if the prediction skill is statistically significant at different lead times. 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

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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.

3 Results

3.1 Prediction skill

Figures 3a and 4a show respectively the correlation and RMSE-SS of the dynamical 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 3b and 4b 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. 3b), with a significant reduction in prediction error beyond ten days for most modes (Fig. 4b). 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 model for atmospheric variables. And the hybrid model exhibits greater improvements in lower-order modes compared to higher-order modes (Fig. 3a and 3b, Fig. 4a and 4b).

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 3c and 4c show the correlation and RMSE-SS of the dynamical model for oceanic temperature and streamfunction. Since the ocean has slower variability than the atmosphere, the dynamical model has significant prediction skills for up to 60 years in oceanic temperature in most modes and oceanic streamfunction in some modes. Overall, odd-numbered modes exhibit higher predictive 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 3d and 4d 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 model, in particular, for oceanic temperature in the first and last modes and oceanic streamfunctions in some modes in which the dynamical model has no prediction skill at all (e.g., $\varphi_{o,2}$ and $\varphi_{o,6}$).

To further demonstrate the advantages of the hybrid model, we use a ten-day lead time for atmospheric variables and a forty-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. 7 for correlation, Fig. 6 and Fig. 8 for RMSE-SS).

For atmospheric variables, both atmospheric streamfunction and temperature exhibit similar spatial characteristics (Figs. 5 and 6). We find that the hybrid model has similar spatial patterns but outperforms the dynamical model in most grid points.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

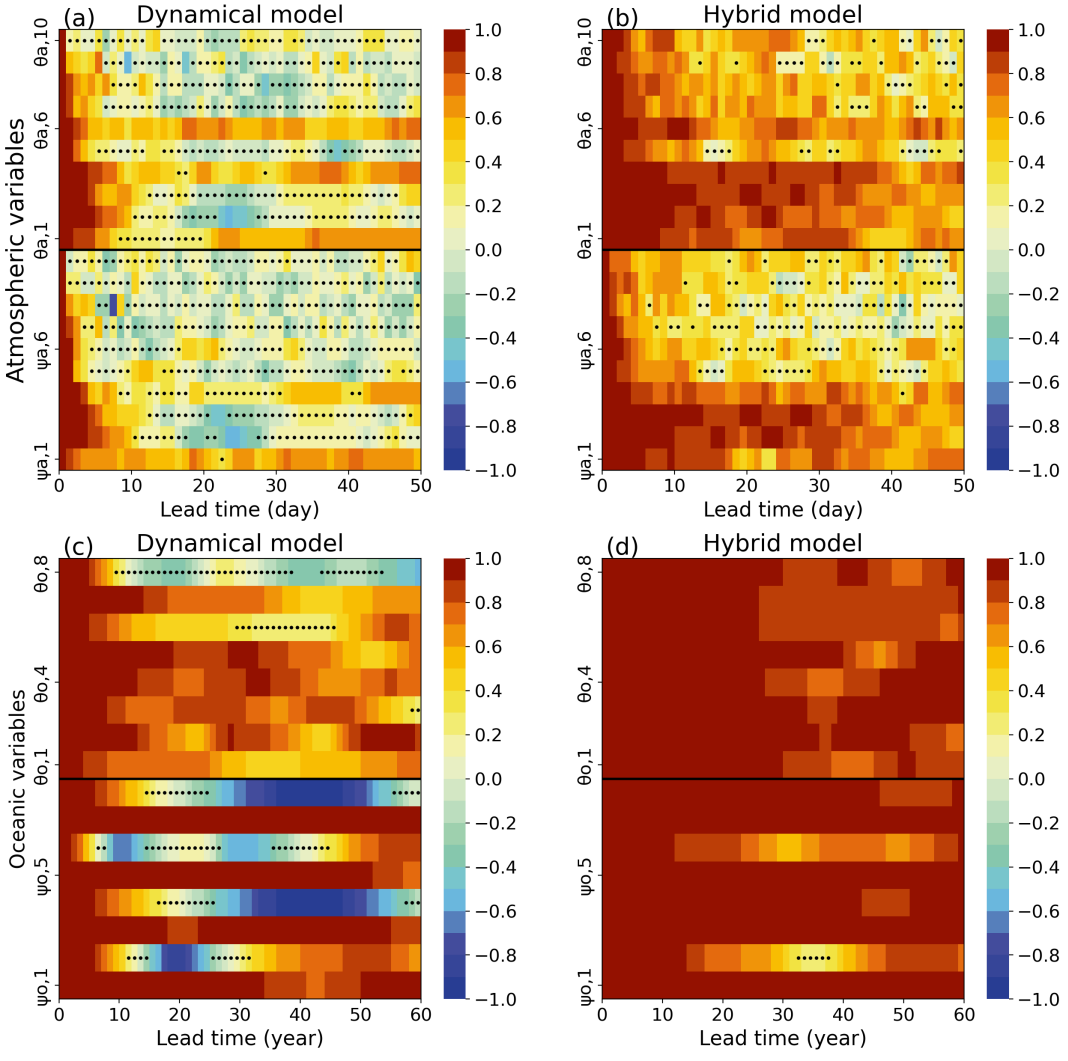

Figure 3. Correlation as a function of the prediction lead time for different variables. (a,c) The correlation between the dynamical model and truth (b,d) The correlation between the hybrid model and truth. The atmospheric variables are calculated based on daily data, while the oceanic variables are based on annual average data. The black dot indicates the correlation does not exceed the 95% significance test.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

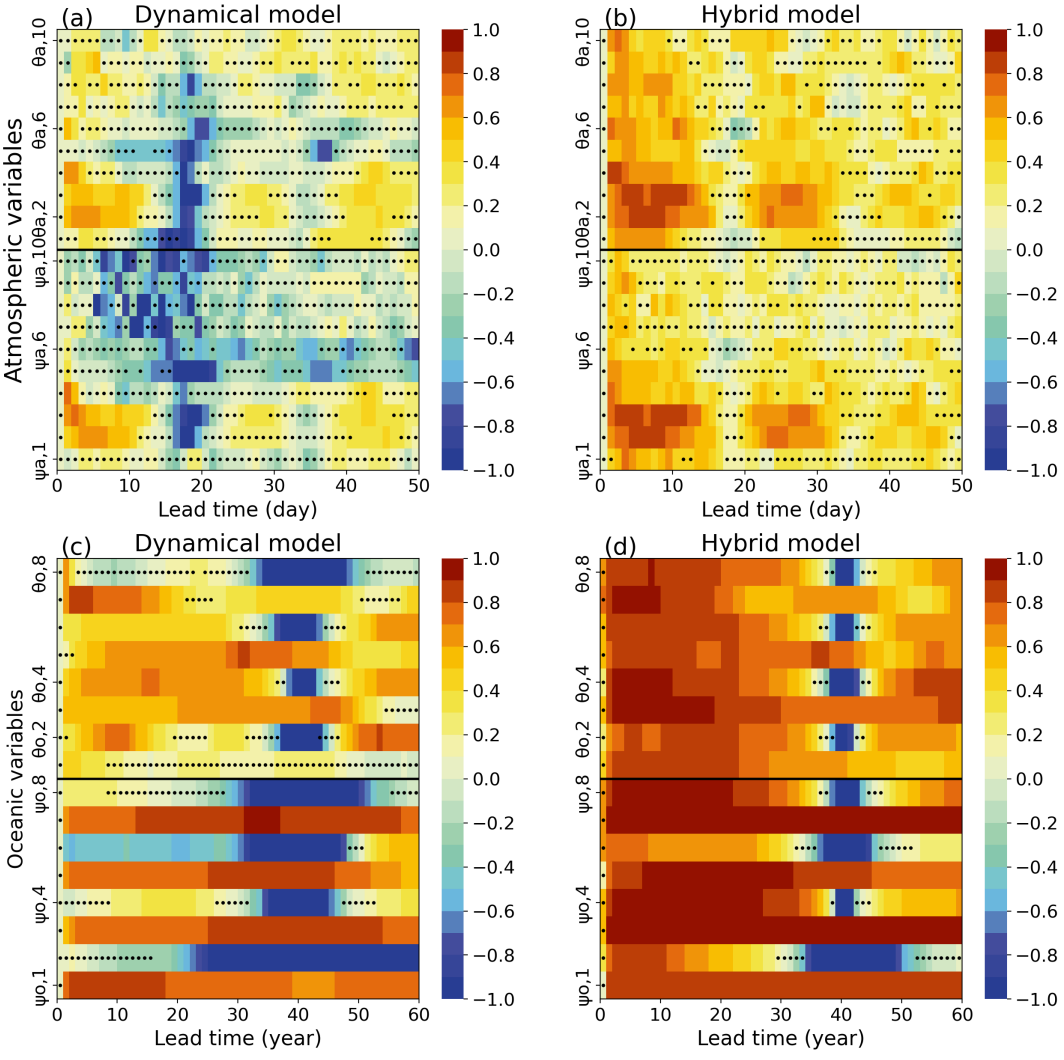

Figure 4. Same as Fig. 3, but for RMSE-SS.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

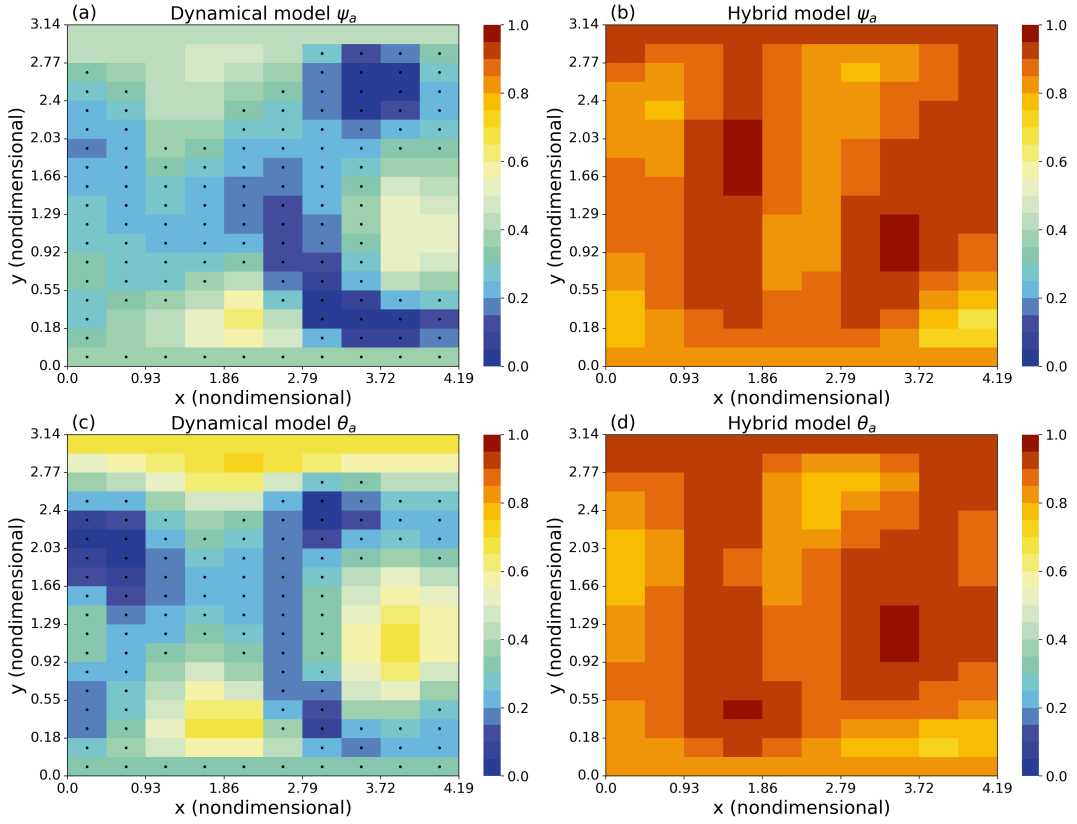

Figure 5. Correlation at the prediction lead day 10 for atmospheric variables. (a,b) The correlation between the dynamical model and the truth for atmospheric temperature, (c,d) The correlation between the dynamical model and the truth for atmospheric streamfunction. The black dot indicates the correlation does not exceed the 95% significance test.

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 grid points (Figures 7a and 8a). Compared to the dynamical model, the hybrid model exhibits higher correlations and RMSE-SS at all grid points, outperforming the dynamical model (Figures 7b and 8d). For oceanic temperature, the dynamical model has higher correlations and significant RMSE-SS in the northwest region. 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 7d and 8d).

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.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

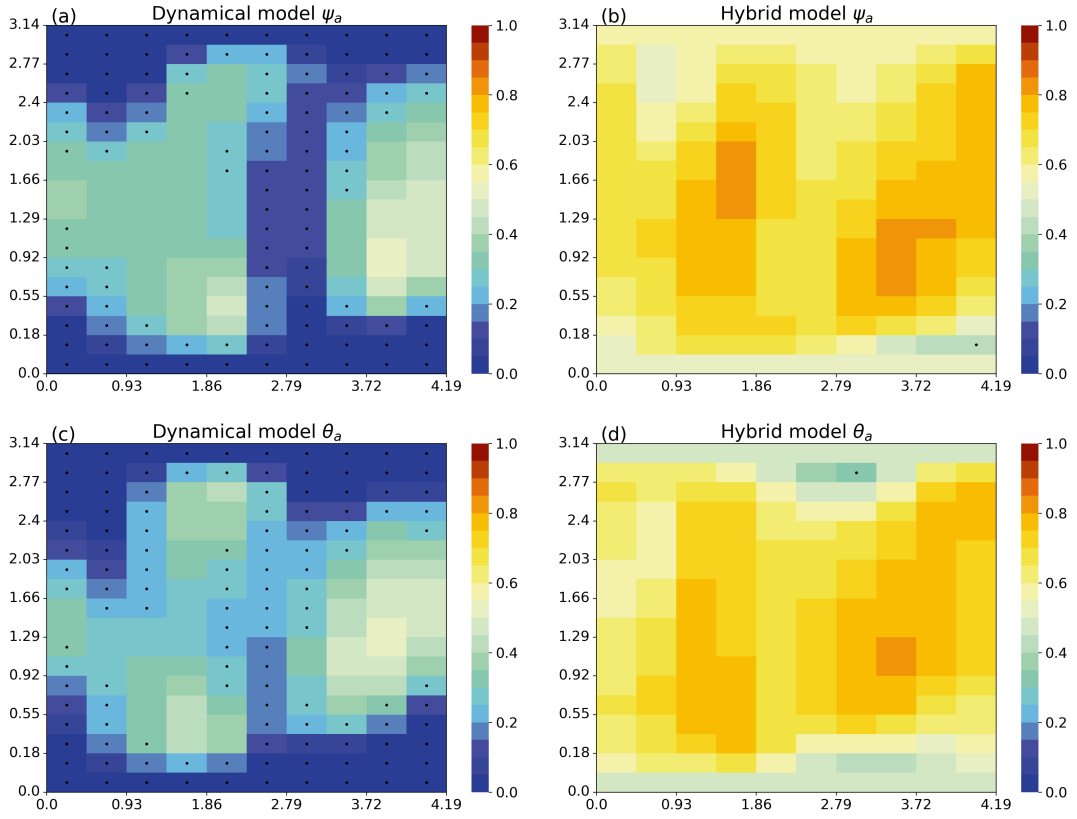

Figure 6. Same as Fig. 5, but for RMSE-SS.

In summary, the overall performance of the hybrid model surpasses that of the dynamical 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 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 into the relative importance of atmospheric and oceanic error correction for the overall prediction performance.

In Figures 9a, 9b, 10a, and 10b, 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

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

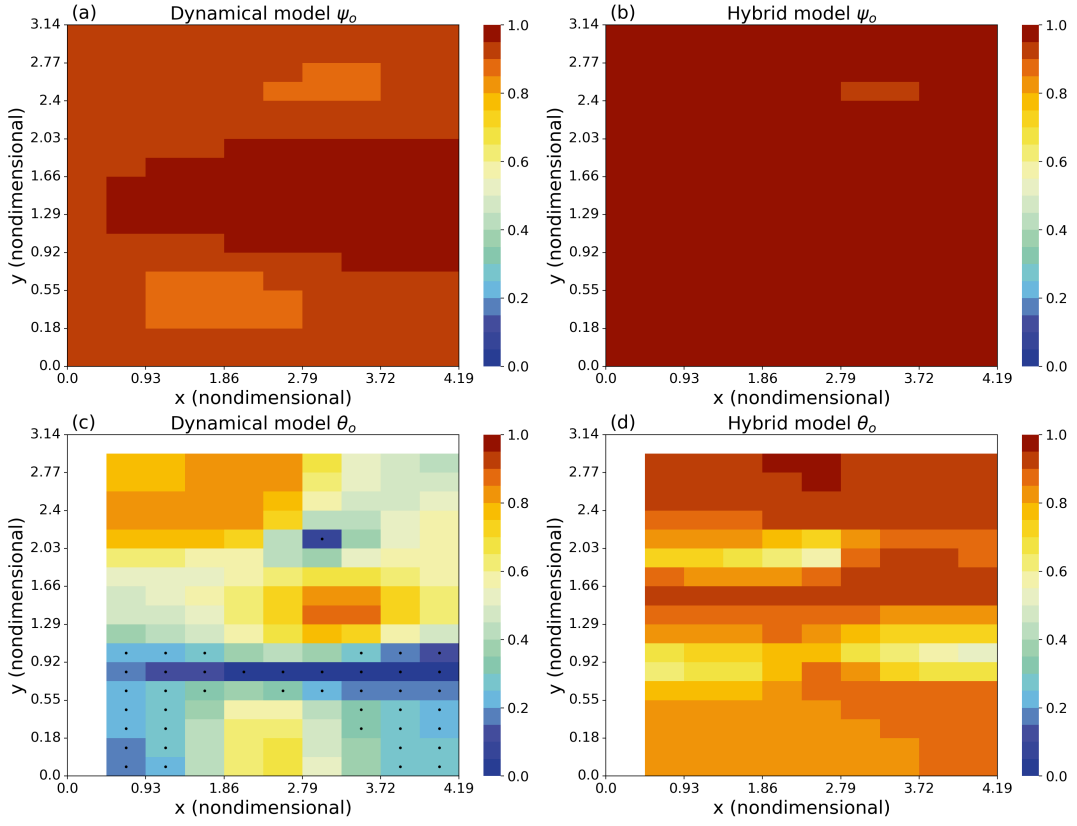

Figure 7. Correlation at the prediction lead year 40 for oceanic variables. (a,b) The correlation between the dynamical model and the truth for oceanic streamfunction, (c,d) the correlation between the dynamical model and the truth for oceanic temperature. The white areas in the temperature map result from consistently zero temperature anomalies at the western and northern boundaries, which prevents the calculation of correlation. The black dot indicates the correlation does not exceed the 95% significance test.

atmospheric errors (green line) and correcting both the atmospheric and oceanic errors (red line). However, in the early forecast 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 model (blue line), we find that correcting only the oceanic errors (cyan line) does not lead to improvements in atmospheric prediction. It is related to the fact in MAOOAM that the atmosphere mostly drives the ocean but the ocean has too weak influences in the atmosphere for short-term climate prediction (Jung and Vitart, 2006).

In Figures 9c, 9d, 10c and 10d, 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 oceanic streamfunction (Fig. 9c), solely correcting oceanic errors (cyan line) does not improve the prediction skill. Specifically, as the lead time increases, it exhibits lower skills compared to the dynamical

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

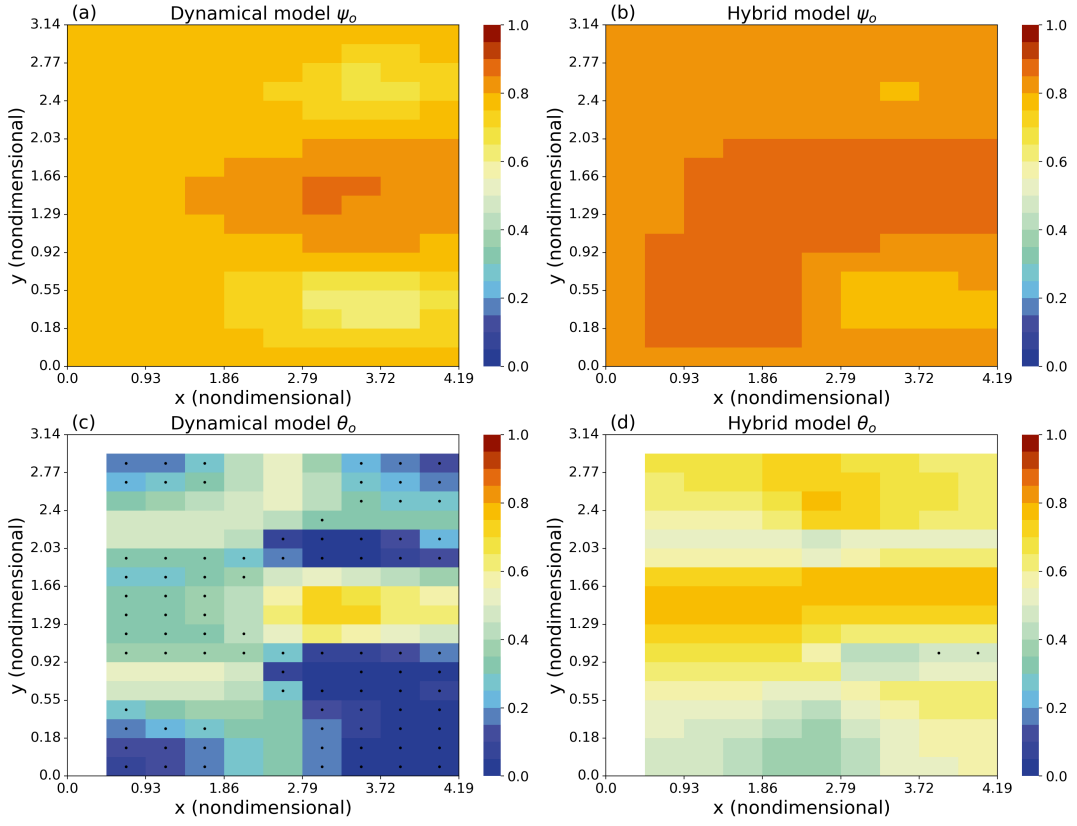

Figure 8. Same as Fig. 7, but for RMSE-SS.

model (blue line). When only correcting atmospheric errors (green line), the improvement in prediction skill occurs in the first 20 years of lead time. However, after 20 years, while the skill of correcting atmospheric errors starts to decline and becomes comparable to the skill of the dynamical model (blue line), simultaneously correcting both atmospheric and oceanic model errors is still better than the dynamical model. For RMSE-SS (Fig. 10c), 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 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 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.

Regarding oceanic temperature (Figs. 9d and 10d), 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.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

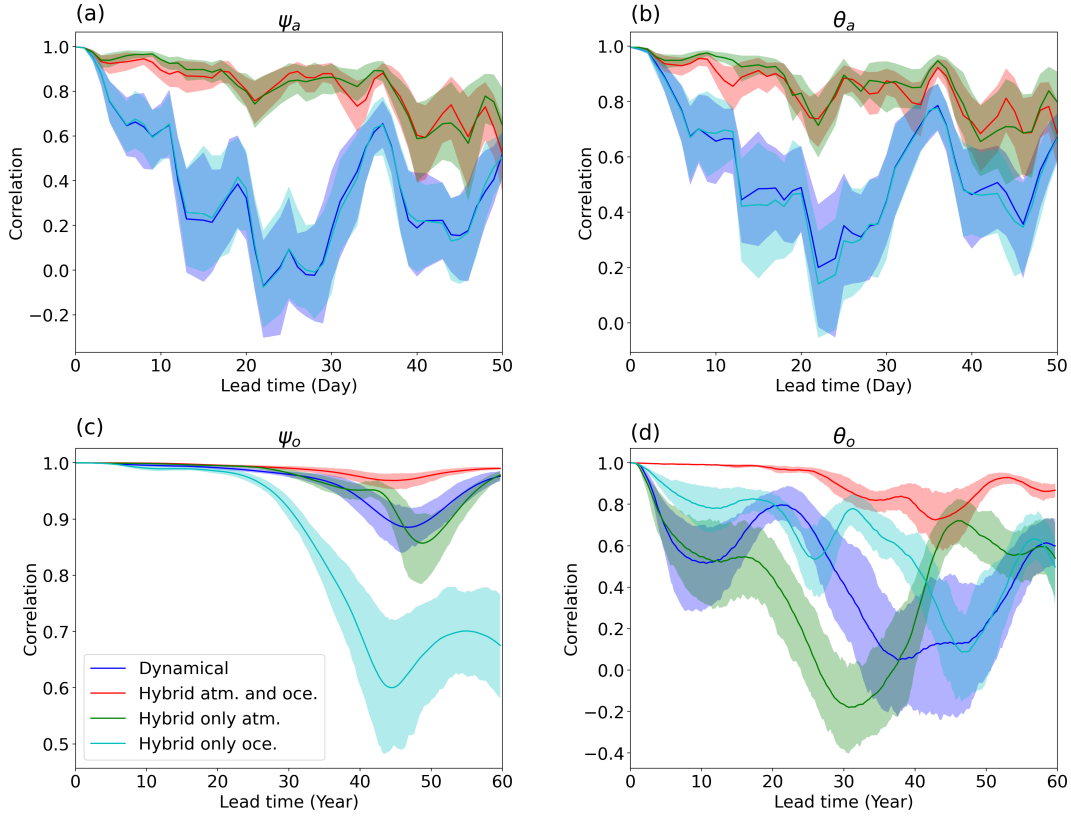

Figure 9. Correlation as a function of lead time (50 days for the atmospheric variable and 60 years for the oceanic variables). Shading shows one standard deviation calculated by the bootstrap method described in section 2.5. The red line is the correlation of the hybrid model built by correcting both atmospheric and oceanic model errors, the green line is the correlation of the hybrid model built by only correcting atmospheric model errors, the cyan line is the correlation of the hybrid model built by only correcting oceanic model errors and the blue line is the correlation of the dynamical model.

To better illustrate the advantages of the hybrid model, we use a set of experimental results as an example to demonstrate the benefits of correcting model errors for long-term simulations (Fig. 11). For atmospheric variables (Fig. 11a and 11b), 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 oceanic streamfunction (Fig. 11c), 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. 11d), 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 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.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

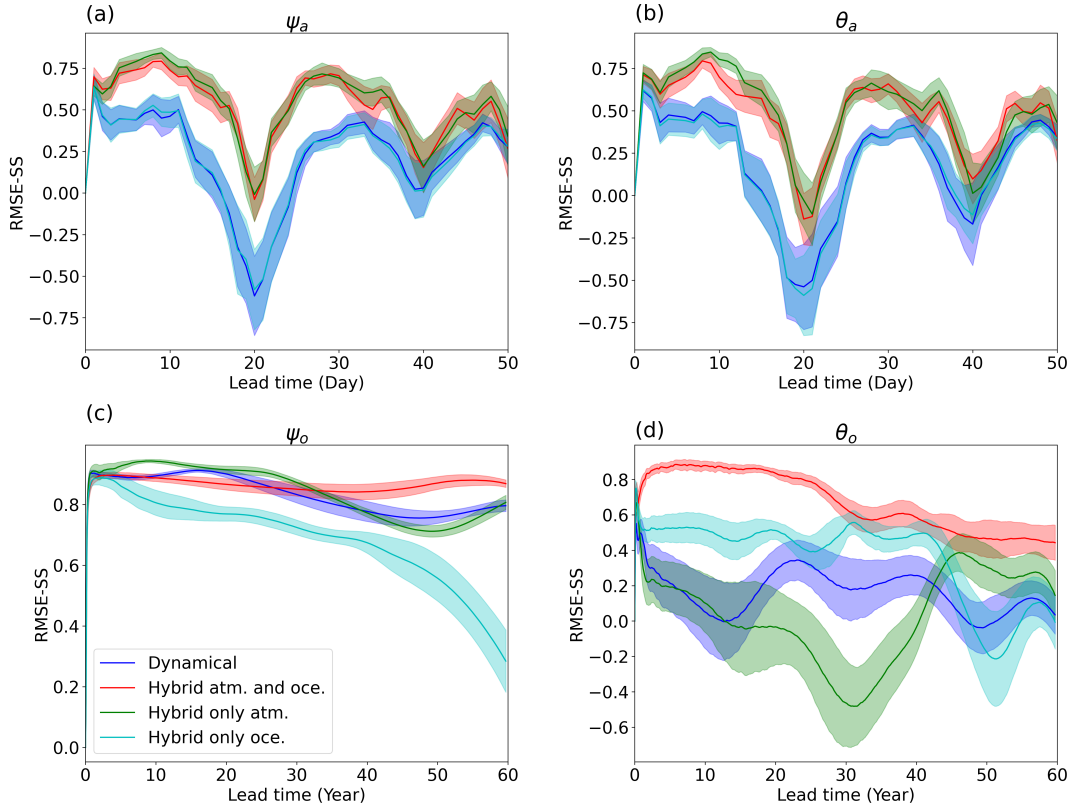

Figure 10. Same as Fig. 9, but for RMSE-SS.

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. We constructed a data-driven predictor of dynamical model error with the ML techniques and integrated it with the dynamical 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 model and achieve more accurate climate predictions.

We also investigated the impact of individually correcting either atmospheric or oceanic model errors on prediction skills. For short-term atmospheric prediction, its accuracy is more influenced by atmospheric errors, while only correcting oceanic errors has little impact on short-term atmospheric prediction Balmaseda and Anderson (2009). For long-term ocean prediction,

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

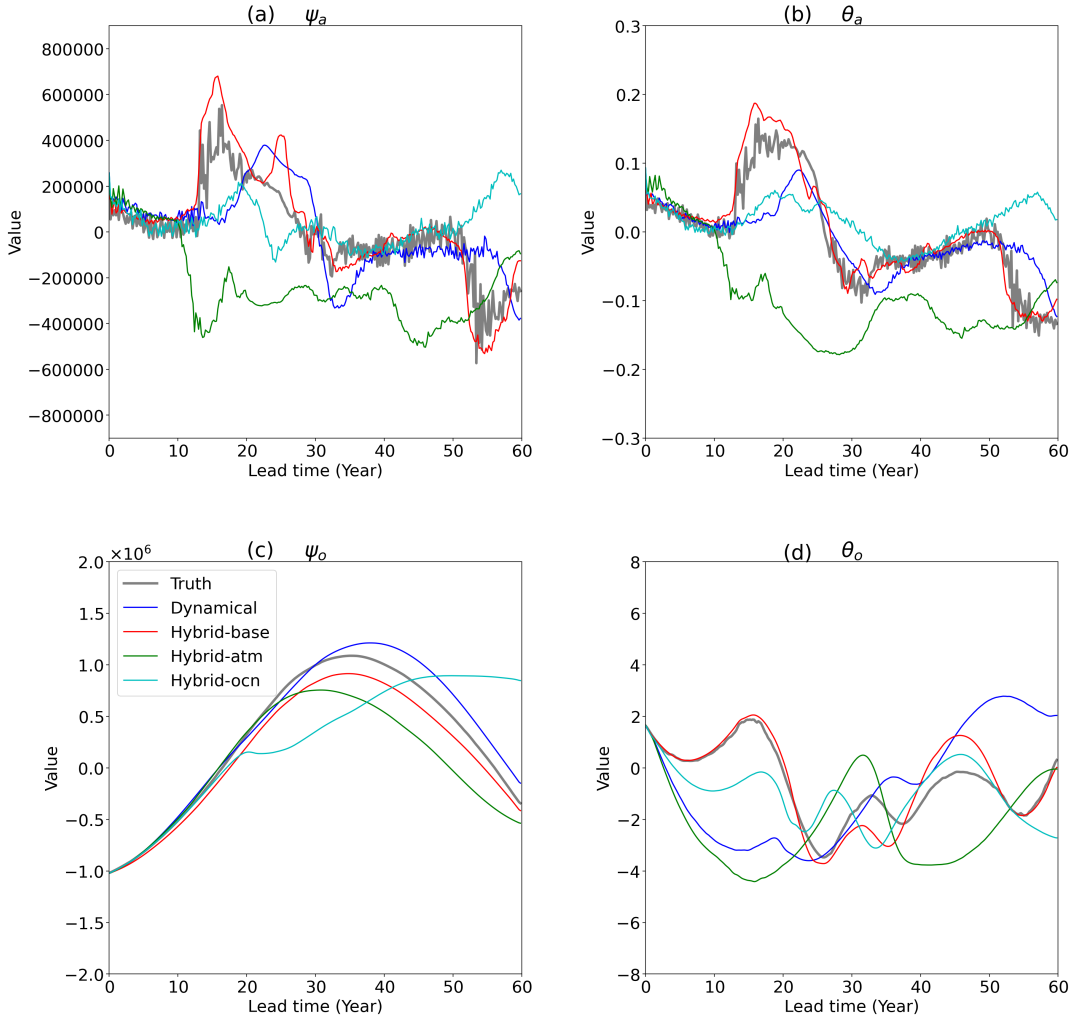

Figure 11. A case study based on the ensemble mean and monthly mean average illustrating the simulation results of four variables averaged over the domain. (a) atmospheric streamfunction, (b) atmospheric temperature, (c) oceanic streamfunction, (d) oceanic temperature.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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 oceanic model errors for long-term climate prediction.

This study serves as a proof of concept, demonstrating the potential of using ML to learn and correct errors in 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), 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).

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

References

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

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, 2021.

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.

https://doi.org/10.5194/egusphere-2025-212
Preprint. Discussion started: 17 February 2025

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.