Enhancing climate predictions with combination of dynamical model and artificial neural network

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

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

[2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China

[3]Nansen Environmental and Remote Sensing Center and Bjerknes Centre for Climate Research, Bergen, Norway

Key Points:

- Artificial neural network (ANN) has the ability to learn errors in a simplified coupled ocean-atmosphere model.
- Combining the ANN-based error correction model with the dynamical model significantly enhanced the prediction skills.
- Correcting both atmospheric and oceanic errors achieved the best prediction skill for climate prediction.

Corresponding author: Xidong Wang, xidong_wang@hhu.edu.cn

Abstract

Dynamical models used in climate prediction often suffer from systematic errors that can deteriorate their predictions. We propose a hybrid model that combines both dynamical model and artificial neural network (ANN) correcting model errors to improve climate predictions. We conducted a series of experiments using the Modular Arbitrary-Order Ocean-Atmosphere Model (MAOOAM) and trained the ANN with input from both atmospheric and oceanic variables and output from analysis increments. Our results demonstrate that the hybrid model outperforms the dynamical model in terms of prediction skill for both atmospheric and oceanic variables across different lead times. Furthermore, we conducted additional experiments to identify the key factors influencing the prediction skill of the hybrid model. We found that correcting both atmospheric and oceanic errors yields the highest prediction skill while correcting only atmospheric or oceanic errors has limited improvement.

Plain Language Summary

Climate prediction is crucial for understanding and preparing for the effects of the atmosphere and the ocean on our societies. However, current climate prediction models (scientific software) can have errors that limit their accuracy. To overcome this, we introduce a hybrid model that combines climate models with the artificial neural network (ANN). The ANN component is trained to identify and correct errors in the climate model. By reducing these errors with ANN, our hybrid model provides more reliable climate predictions. This is important for decision-making and planning related to climate impacts.

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 models, such as ocean-atmosphere coupled general circulation models, have been widely used for climate predictions (e.g., F. J. Doblas-Reyes et al., 2013; 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 & Anderson, 2009; F. Doblas-Reyes et al., 2013) have been evolving towards the use of data assimilation (DA, Carrassi et al., 2018) which combines observations with dynamical models to best estimate the state of the climate system (S. G. Penny & Hamill, 2017). Reducing the model error is challenging since the model error can be caused by many factors, e.g., model parameterizations (e.g., T. N. Palmer, 2001), unresolved physical processes (e.g., Moufouma-Okia & Jones, 2015), or numerical approximations (e.g., Williamson et al., 1992). Although there have been massive efforts in climate model development, the model error remains significantly large (e.g., Richter, 2015; T. Palmer & Stevens, 2019; Richter & Tokinaga, 2020; Tian & Dong, 2020).

There is a growing interest in utilizing machine learning (ML) techniques to address errors in dynamical models. 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).

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 prior to the numerical integration of the model. In a similar vein, 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 ocean-atmosphere model called the Modu-

lar Arbitrary-Order Ocean-Atmosphere Model (MAOOAM) (De Cruz et al., 2016). Their study aimed to infer model errors associated with unresolved processes within the dynamical model.

Several other investigations (e.g., 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, the availability of observational data for training, validation, and testing is relatively limited. These factors impose limitations on exploring the full potential of developing a data-driven predictor for model errors.

Furthermore, in the works mentioned here-before, the hybrid model is tested in an idealized setting in which initial conditions are perfectly known. In realistic climate predictions, there is uncertainty in initial conditions which is generally represented as an ensemble of initial conditions, and an ensemble of predictions is obtained (Wang et al., 2019). To our knowledge, the skill of hybrid models in the realistic case of imperfect initial conditions with an ensemble of forecasts has not been thoroughly assessed.

In this study, we aim at filling this gap. We utilize the low-order coupled ocean-atmosphere model named 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 a function of lead time. Furthermore, we aim to identify whether atmosphere model errors or ocean model errors play a pivotal role in degrading climate prediction accuracy. This study presents novel findings as it directly addresses a research gap in our current understanding. The insights obtained from this research hold significant value for the climate prediction community, contributing to advancements in the field.

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 simplify the analysis by considering model errors solely attributed to coarse resolutions. We adopt similar configurations of the model (section 2.1), DA technique (section 2.2), and Artificial Neural Network (ANN) approach (section 2.3) as outlined by Brajard et al. (2021). However, it is important to note that our objectives are different. While they focused on methodological developments, our primary aim is to investigate how the benefits of ML-based error correction evolve with lead time for climate prediction purposes. Furthermore, our experimental setup incorporates a more realistic approach. For further details, please refer to 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 replicate climate variability. MAOOAM has been widely employed in qualitative analyses for various purposes (e.g., S. 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 stream function ψ_a and temperature anomaly θ_a, as well as n_o ($n_o = d_{oy}d_{ox}$) modes of the oceanic stream function 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 $2n_a + 2n_o$. 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, while the ocean operates at a 2x-4y (i.e., $d_{ox} = 2$ and $d_{oy} = 4$) resolution.

On the other hand, the **M36** configuration consists of 36 variables, with 10 atmospheric modes ($n_a = 10$) and the same 8 oceanic modes ($n_o = 8$) as in **M56**. The atmospheric component in **M36** operates at a 2x-2y (i.e., $d_{ax} = 2$ and $d_{ay} = 2$) resolution, while the ocean component maintains a 2x-4y (i.e., $d_{ox} = 2$ and $d_{oy} = 4$) resolution, identical to that of **M56**.

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 less than **M56** in the y-direction. This difference leads to a lack of higher-order atmospheric modes in **M36**, thereby resulting in an inability to capture variability on small scales. Consequently, the primary source of model error in this study is attributed to the coarse resolution of the model.

2.2 Ensemble Kalman Filter

The 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 constructing the background error covariance from the dynamical ensemble is more reliable than other DA methods using the static error covariance (e.g., Sakov & Sandery, 2015). Moreover, the utilization of an ensemble-based error covariance ensures that the assimilation updates adhere 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 Figure 1. Specifically, we employ the finite-size ensemble Kalman filter (EnKF-N) method proposed by Bocquet et al. (2015). This method automatically estimates the inflation factor, a critical parameter in ensemble DA systems, thereby enhancing the performance of the assimilation experiments.

It is worth mentioning that we expect no significant alterations in the conclusions of this paper when using the traditional EnKF instead of EnKF-N. Hence, for simplicity, we refer to both methods as the EnKF in the following discussions, as their differences do not have a substantial impact on the overall outcomes of this study.

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 truth state at time t_{k+1}.

We aim to use ANN to emulate the model error ε. For simplicity, our ANN configuration is set to the same as that of Brajard et al. (2021). 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 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 the ANN and ε'_{k+1} represents the model error estimated by the ANN. The full state at time t_{k+1} of the hybrid model can be expressed as follows:

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

2.4 Experimental settings

We present the experimental setup in Figure 1. The experiments are conducted using two configurations of MAOOAM, as described in section 2.1. The configuration with 56 variables (referred to as **M56**, section 2.1) represents the true climate system, while the configuration with 36 variables (referred to as **M36**) represents a dynamical prediction system. The experiments (depicted in Figure 1) 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 219 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 after subtracting the one-month running average ($\boldsymbol{\sigma}^{\mathrm{hf}}$). Observations are generated at intervals of approximately 27 hours.
- We perform a simulation with 50 ensemble members. The initial conditions of the ensemble are randomly sampled from a long free-run simulation of **M36** after the spin-up period. We assimilate synthetic observations and generate an analysis dataset with an ensemble size of 50.

- We generate two sets of ensemble predictions, each consisting of 50 members. The first set is based on the dynamical model (**M36**), while the second set is based on the hybrid model. The prediction experiments start in each second year from the year 125 to the year 185, with each experiment lasting for 30 years. Each prediction consists of 50 ensemble members. The initial conditions for these ensembles are taken from the analysis conducted (refer to Figure 1).

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 (Figure 1).
- Validation/testing data: The latter 94.6 years of the dataset are used to validate the ANN training and initialize prediction experiments (Figure 1).

It is worth noting that we employ the identical ANN configurations as outlined in Brajard et al. (2021) who have developed the methodology in MAOOAM. In this study, the ANN parameters are trained only once, without any modifications to the ANN model throughout the training process. We examined the loss curves (Figure S1) to assess the suitability of the ANN model for our specific application. The training curves provided evidence that the network was continuing to learn throughout the training process. To simplify, we utilize the same dataset for both validation and testing purposes.

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 20 days for atmospheric variables and up to 30 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 their capabilities and limitations. To do so, our experimental setup is different 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 the dataset into two distinct parts: one for training the

ANN and the other for validation/test 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) using 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 forecasts 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, we employ the root mean square skill score (RMSE-SS), a commonly used metric in weather forecasting and climate prediction. The RMSE-SS compares the root mean square error (RMSE) of the prediction to the RMSE of a persistence prediction. It is defined as:

$$\mathrm{RMSE-SS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{prediction}}}{\mathrm{RMSE}_{\mathrm{persistence}}}, \tag{6}$$

where $\mathrm{RMSE}_{\mathrm{prediction}}$ represents the RMSE between the prediction (ensemble mean) and the corresponding 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 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 Figure 2.

To assess the significance of the RMSE-SS results, we employ a two-tailed Student's t-test to compare the mean squared errors of the prediction and persistence. This statistical test helps determine if the difference between the two sets of errors is statistically significant. To estimate the uncertainties of the RMSE-SS, we utilize the bootstrap method. We randomly select, with replacement, 30 data points from the 30 prediction experiments and calculate the RMSE-SS based on this sampled data. This procedure is repeated 10,000 times, resulting in a sample of 10,000 RMSE-SS values. The standard deviation of this sample is then used to estimate the uncertainties associated with the 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 RMSE-SS values obtained from the prediction experiments.

3 Result

3.1 Prediction skill

Figure 2a presents the prediction skills of the dynamical model for atmospheric temperature (θ_a) and stream function (ψ_a). Notably, the variables associated with lower-order atmospheric modes, such as $\psi_{a,2}$, $\psi_{a,3}$, $\theta_{a,2}$, and $\theta_{a,3}$, exhibit significant prediction skills for up to 14 days. On the other hand, the temperature in higher-order modes demonstrates significant prediction skills within an 8-day lead time, while the stream function in higher-order modes shows no prediction skill throughout the forecast period.

Figure 2b shows the prediction skills of the hybrid model for atmospheric variables. Regarding temperature, the hybrid model exhibits skillful predictions for up to 18 days across most modes. For the stream function, the hybrid model demonstrates skillful predictions for lower-order atmospheric modes for up to 20 days and for higher-order modes for up to 14 days (with the exception of $\psi_{a,9}$, which extends up to 20 days). Overall, the hybrid model outperforms the dynamical model significantly in terms of prediction skills for atmospheric variables.

Figure 2c illustrates the prediction skills of the dynamical model for oceanic temperature and stream function. Due to the lower variability of the ocean compared to the atmosphere, the dynamical model displays significant prediction skills

for oceanic temperature for up to 30 years in most modes, as well as for oceanic stream function in certain modes. Notably, temperature exhibits higher predictability compared to the stream function. However, the ocean stream function variables with even numbers exhibit a lack of skill, which may be attributed to the discrepancy in resolution in the y-direction between the true model's atmosphere and the dynamical model.

In Figure 2d, the prediction skills of the hybrid model are presented. The hybrid model demonstrates significant prediction skills for both oceanic temperature and stream function across all modes for up to 30 years. It is worth mentioning that the hybrid model yields higher RMSE-SS values compared to the dynamical model, particularly for oceanic temperature in the first and last modes, as well as for certain oceanic stream functions where the dynamical model shows no prediction skill at all (e.g., $\psi_{o,2}$ and $\psi_{o,6}$).

In Movies S1-S4, we provide examples of restoring variables in the physical space. These examples highlight that the hybrid model exhibits a closer behavior to the truth in terms of spatial distribution and temporal evolution compared to the dynamical model. 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 30-year prediction period.

The overall performance of the hybrid model surpasses that of the dynamical model, demonstrating the advantages of incorporating a data-driven error correction model constructed by the ANN. This highlights the potential benefits of leveraging data-driven approaches to improve climate predictions.

3.2 Sensitivity experiments

In this section, we extend our analysis by constructing two additional hybrid models to assess the importance 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 only atmospheric errors or only oceanic errors. By comparing the prediction skills of the three key variables of MAOOAM (Vannitsem,

2015) - $\psi_{a,1}$, $\psi_{o,2}$, and $\theta_{o,2}$ - in these hybrid models, we aim to identify which component's error correction has a greater impact on the predictions. Through this analysis, we gain insights into the relative importance of atmospheric and oceanic error correction for the overall prediction performance.

In Figure 3a, we present the prediction skill of different models specifically for the key atmospheric variable $\psi_{a,1}$. Please refer to Figures S2 and S3 for the prediction skill of other atmospheric variables. We observe that there is minimal difference in prediction skill between correcting only the atmospheric errors (purple line) and correcting both the atmospheric and oceanic errors (cyan line). When comparing the hybrid models with the dynamical model result (black dashed line), we find that correcting only the oceanic errors (blue line) does not lead to improvements in atmospheric prediction within a 20-day lead time. This suggests that the influence of low-frequency variability originating from the ocean has a limited impact on short-term predictions of atmospheric variables. These findings indicate that atmospheric error correction plays a more significant role in improving the short-term prediction skill of $\psi_{a,1}$, while the oceanic error correction alone does not provide noticeable benefits in this context.

In Figures 3b and 3c, we focus on the prediction skill of various hybrid models for the two crucial oceanic variables, $\psi_{o,2}$ and $\theta_{o,2}$. Please refer to Figures S4 and S5 for prediction skill analysis of other oceanic variables. Our results reveal that the highest prediction skill over a 30-year period is achieved when both atmospheric and oceanic errors are corrected (cyan line). For $\psi_{o,2}$, correcting solely oceanic errors (blue line) yields a minor enhancement. When correcting atmospheric errors (purple line), a noticeable improvement in prediction skill occurs except for the first five years. This phenomenon may stem from the time required for the oceanic processes to adjust to an error-corrected atmosphere. Regarding $\theta_{o,2}$, we note that either correcting oceanic errors (blue line) or atmospheric errors (purple line) results in positive RMSE-SS values. This suggests that both forms of error correction contribute to enhancing the prediction skill of $\theta_{o,2}$. Furthermore, these two hybrid models significantly outperform the results of the dynamical model (black dashed line) from lead year 15 to lead year 25. In summary, the correction of both atmospheric and oceanic errors proves to be more effective in enhancing prediction accu-

racy for oceanic variables compared to addressing just one component. Noteworthy improvements are observable, particularly after a few years of prediction.

4 Conclusions and Discussions

In this study, we applied a method to online correct the error in a simplified ocean-atmosphere coupled model (MAOOAM). We constructed a data-driven predictor of 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 of both atmospheric and oceanic variables at different lead times. This approach allowed us to mitigate the limitations of the dynamical model and achieve more accurate climate predictions.

We also investigated the impact of correcting either atmospheric or oceanic model errors individually. Our findings revealed that correcting both atmospheric and oceanic errors achieved the best prediction skill for short-term atmosphere prediction and long-term ocean prediction. Correcting only oceanic errors showed some improvement in long-term ocean prediction but a very limited effect on short-term atmosphere prediction. On the other hand, correcting only atmospheric errors effectively improved short-term atmosphere prediction and slightly enhanced long-term ocean prediction.

This study serves as a proof of concept, demonstrating the potential of using ML to learn and correct climate model errors, thus enhancing the prediction skills of climate models. Future applications could involve applying this method to realistic climate models, which are inherently more complex than MAOOAM, and exploring the prediction skills under such conditions.

Open Research Section

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`.

Acknowledgments

This 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 (422003165), 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 Research 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).

References

Balmaseda, M., & Anderson, D. (2009). Impact of initialization strategies and observations on seasonal forecast skill. *Geophysical research letters*, *36*(1).

Bocquet, M., Raanes, P. N., & Hannart, A. (2015). Expanding the validity of the ensemble kalman filter without the intrinsic need for inflation. *Nonlinear Processes in Geophysics*, *22*(6), 645–662.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., ... others (2016). The decadal climate prediction project (dcpp) contribution to cmip6. *Geoscientific Model Development*, *9*(10), 3751–3777.

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2021). Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200086.

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., ... Harris, L. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794.

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, *9*(5), e535.

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., & Tulich, S. (2022). Correcting systematic and state-dependent errors in the noaa fv3-

gfs using neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(11).

De Cruz, L., Demaeyer, J., & Vannitsem, S. (2016). The modular arbitrary-order ocean-atmosphere model: Maooam v1. 0. *Geoscientific Model Development*, *9*(8), 2793–2808.

Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., ... Van Oldenborgh, G. (2013). Initialized near-term regional climate change prediction. *Nature communications*, *4*(1), 1715.

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

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

Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., & Malartic, Q. (2021). A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *Journal of computational science*, *55*, 101468.

Hinton, G., Srivastava, N., & Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, *14*(8), 2.

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

Karspeck, A. R., Yeager, S., Danabasoglu, G., Hoar, T., Collins, N., Raeder, K., ... Tribbia, J. (2013). An ensemble adjustment kalman filter for the ccsm4 ocean component. *Journal of Climate*, *26*(19), 7392–7413.

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A., ... others (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896.

Moufouma-Okia, W., & Jones, R. (2015). Resolution dependence in simulating the african hydroclimate with the hadgem3-ra regional climate model. *Climate Dynamics*, *44*(3-4), 609–632.

Palmer, T., & Stevens, B. (2019). The scientific challenge of understanding and estimating climate change. *Proceedings of the National Academy of Sciences*, *116*(49), 24390–24395.

Palmer, T. N. (2001). 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*(572), 279–304.

Penny, S., Bach, E., Bhargava, K., Chang, C.-C., Da, C., Sun, L., & Yoshida, T. (2019). Strongly coupled data assimilation in multiscale media: Experiments using a quasi-geostrophic coupled model. *Journal of Advances in Modeling Earth Systems*, *11*(6), 1803–1829.

Penny, S. G., & Hamill, T. M. (2017). Coupled data assimilation for integrated earth system analysis and prediction. *Bulletin of the American Meteorological Society*, *98*(7), ES169–ES172.

Raanes, P. N. (2018, December). *nansencenter/dapper: Version 0.8.* Retrieved from `https://doi.org/10.5281/zenodo.2029296` doi: 10.5281/zenodo.2029296

Richter, I. (2015). Climate model biases in the eastern tropical oceans: Causes, impacts and ways forward. *Wiley Interdisciplinary Reviews: Climate Change*, *6*(3), 345–358.

Richter, I., & Tokinaga, H. (2020). An overview of the performance of cmip6 models in the tropical atlantic: mean state, variability, and remote impacts. *Climate Dynamics*, *55*(9-10), 2579–2601.

Sakov, P., & Sandery, P. A. (2015). Comparison of enoi and enkf regional ocean reanalysis systems. *Ocean Modelling*, *89*, 45–60.

Tian, B., & Dong, X. (2020). The double-itcz bias in cmip3, cmip5, and cmip6 models based on annual mean precipitation. *Geophysical Research Letters*, *47*(8), e2020GL087232.

Vannitsem, S. (2015). The role of the ocean mixed layer on the development of the north atlantic oscillation: A dynamical system's perspective. *Geophysical Research Letters*, *42*(20), 8615–8623.

Wang, Y., Counillon, F., Keenlyside, N., Svendsen, L., Gleixner, S., Kimmritz, M., ... Gao, Y. (2019). Seasonal predictions initialised by assimilating sea surface temperature observations with the enkf. *Climate Dynamics*, *53*, 5777–5797.

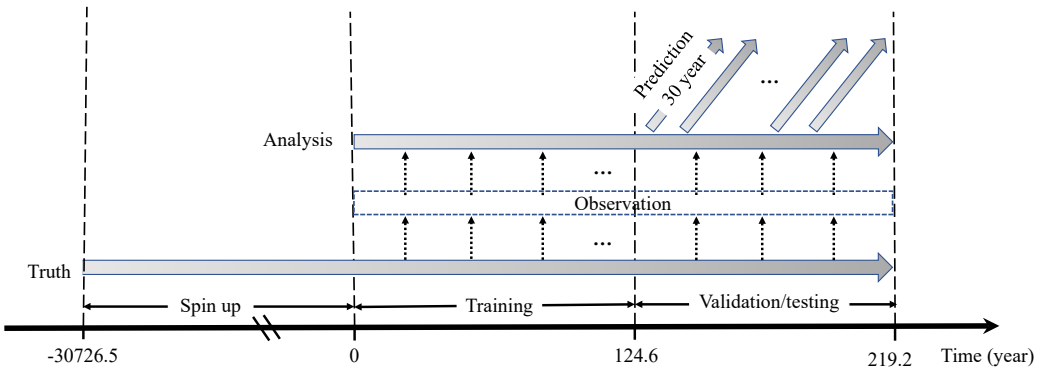

Figure 1. Schematic of experiments listed in section 2.4.

Watson, P. A. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, *11*(5), 1402–1417.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555.

Williamson, D. L., Drake, J. B., Hack, J. J., Jakob, R., & Swarztrauber, P. N. (1992). A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of computational physics*, *102*(1), 211–224.

Zhang, S., Harrison, M., Rosati, A., & Wittenberg, A. (2007). System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Monthly Weather Review*, *135*(10), 3541–3564.

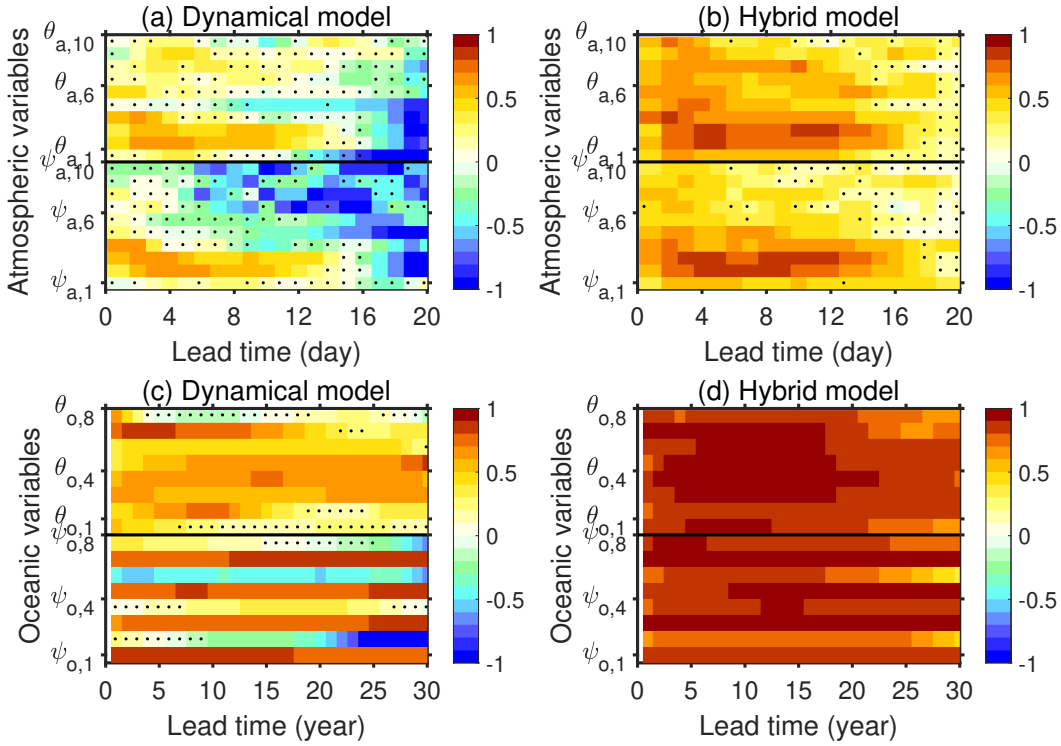

Figure 2. RMSE-SS as a function of the prediction lead time for different variables. (a,c) The RMSE-SS of the dynamical model (b,d) the RMSE-SS of the hybrid model. The black dot indicates the RMSE-SS does not exceed the 95 significance test.

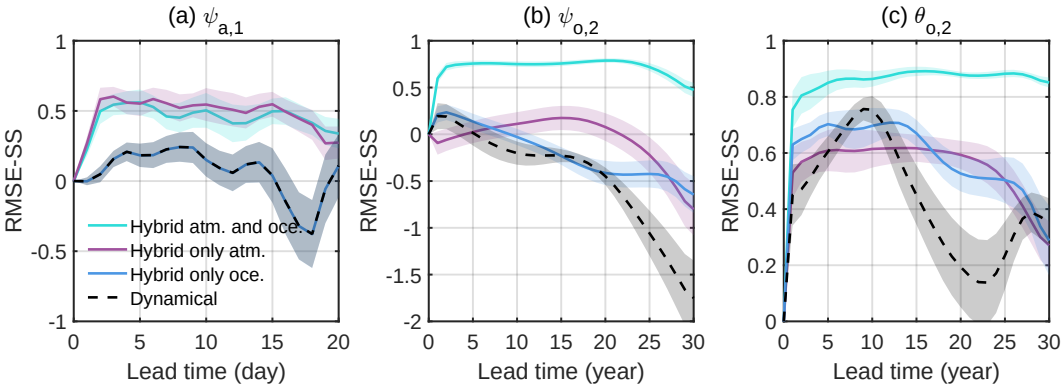

Figure 3. RMSE-SS for three key variables (a) $\psi_{a,1}$, (b) $\psi_{o,2}$ and (c) $\theta_{o,2}$ as a function of lead time (20 days for the atmospheric variable and 30 years for the ocean variables). The cyan line represents the RMSE-SS of the hybrid model that corrects both atmospheric and oceanic model errors. The purple line corresponds to the RMSE-SS of the hybrid model that corrects only atmospheric model errors, while the blue line represents the RMSE-SS of the hybrid model that corrects only oceanic model errors. The dashed black line represents the RMSE-SS of the dynamical model. The shading represents one standard deviation calculated using the bootstrap method described in section 2.5. The shaded area provides an estimate of the uncertainty associated with the RMSE-SS values.