Improve dynamical climate prediction with machine learning

Zikang He[1], Julien Brajard[2], Yiguo Wang[3], Xidong Wang[1], and Zheqi Shen[4]

[1]College of Oceanography, Hohai University
[2]NERSC
[3]Nersc
[4]Hohai University

March 16, 2023

Abstract

Dynamical models used in climate prediction often have systematic errors that can deteriorate predictions. In this study, we work in a twin experiment framework with a reduced-order coupled ocean-atmosphere model and aim to demonstrate the benefit of machine learning for climate prediction. Machine learning is applied to learn the model error and thus build a data-driven model to emulate the dynamical model error. Then we build a hybrid model by combining the data-driven and dynamical models. The prediction skill of the hybrid model is compared to that of the standalone dynamical model. We applied this approach to the ocean-atmosphere coupled model. The results show that the hybrid model outperforms the dynamical model alone for both atmospheric and oceanic variables. Also, we build two other hybrid models only correcting either atmospheric errors or oceanic errors. It was found that correcting both atmospheric and oceanic errors leads to the best performance.

Improve dynamical climate prediction with machine learning

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 can learn the error of a simplified coupled ocean-atmosphere model.
- The hybrid model combining the artificial neural network and the dynamical model shows good performance to improve dynamic prediction skills.
- The hybrid model overperforms the dynamical model for both atmospheric and oceanic variables.

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

Abstract

Dynamical models used in climate prediction often have systematic errors that can deteriorate predictions. In this study, we work in a twin experiment framework with a reduced-order coupled ocean-atmosphere model and aim to demonstrate the benefit of machine learning for climate prediction. Machine learning is applied to learn the model error and thus build a data-driven model to emulate the dynamical model error. Then we build a hybrid model by combining the data-driven and dynamical models. The prediction skill of the hybrid model is compared to that of the standalone dynamical model. We applied this approach to the ocean-atmosphere coupled model. The results show that the hybrid model outperforms the dynamical model alone for both atmospheric and oceanic variables. Also, we build two other hybrid models only correcting either atmospheric errors or oceanic errors. It was found that correcting both atmospheric and oceanic errors leads to the best performance.

Plain Language Summary

Dynamical models are essential for predicting the climate and for studying the Earth's system. But they still have some errors that cannot be corrected. Recently, a lot of progress has been made in machine learning methods based on the large quantities of observations collected. These are data-driven algorithms that learn from existing data. We show the possibility that applying machine learning to a simplified ocean-atmospheric coupled model. After being presented with enough data from the climate model, the network can successfully predict the model's error, thus correcting the error of the dynamical model. This finding provides an idea for error correction in coupled models and is important for real applications.

1 Introduction

Dynamical models, such as ocean-atmosphere coupled general circulation models, have been widely used for climate predictions over the past few decades, e.g., seasonal predictions (F. J. Doblas-Reyes et al., 2013) and decadal predictions (Boer et al., 2016) (DCPP). 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 (Balmaseda & Anderson, 2009; F. Doblas-Reyes et al., 2013), most prediction centers 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 (Penny & Hamill, 2017). Meanwhile, although there have been massive efforts in model development, the model error remains significantly large (Palmer & Stevens, 2019; Tian & Dong, 2020). It is because many factors (e.g., unknown physical law, unresolved small-scale processes, and numerical integration errors) can cause the model error (Hawkins & Sutton, 2009).

Machine learning (ML) can efficiently extract useful information from data (Salcedo-Sanz et al., 2020). It has been used to build a data-driven predictor of the model error which is combined with a dynamical model to produce a statistical-dynamical hybrid model (Watson, 2019; Farchi et al., 2021; Brajard et al., 2021). Watson (2019) worked in a low-order Lorenz model and applied ML to correct the error from time step to time step. They found that the approach maintained the model stable and improved predictions. Farchi et al. (2021) worked in the two-scale Lorenz model and compared the error corrections added as an extra term (i.e., resolvent correction) or directly inside the tendencies of the dynamical model (i.e., tendency correction). They showed that the tendency correction performed better but was more technical than the resolvent correction. Brajard et al. (2021) applied ML into the two-scale Lorenz model and a low-order coupled ocean-atmosphere model called Modular Arbitrary-Order Ocean-Atmosphere Model (MAOOAM) (De Cruz et al., 2016) to infer the model error related to unresolved processes from the state of the dynamical model. Brajard et al. (2021) mostly focused on presenting and validating their methodology and barely presented the prediction improvements for atmospheric variables at a one-day lead time and oceanic variables at a two-year lead time. However, they did not investigate how the improvement evolves as a function of lead time and how long the improvement remains significant. In addition, Brajard et al. (2021) used perfect initial conditions in prediction experiments, which is not a realistic setting because initial conditions are never perfectly known in reality.

In this study, we set up a more realistic framework than Brajard et al. (2021) and aim to explore the potential of ML-based model error correction for climate prediction at different lead times, which is valuable for climate prediction communities.

We also aim to identify whether errors in the atmosphere or the ocean play a key role in degrading prediction skills.

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 make use of MAOOAM (De Cruz et al., 2016) which is able to mimic climate variability and is numerically cheap to perform a large number of experiments. We employed the same configurations of the model (section 2.1), DA (section 2.2) and Artificial Neural Network (ANN, section 2.3) as Brajard et al. (2021). However, our experiments are more realistic and different from that of Brajard et al. (2021). Please refer to section 2.4 for details.

2.1 Modular Arbitrary-Order Ocean-Atmosphere Model

MAOOAM consists of a two-layer quasi-geostrophic (QG) atmospheric component coupled both thermally and mechanically to a QG shallow-water oceanic component. The coupling between the two components includes wind forcings, radiative and heat exchanges. The model variables are described in the spectral modes. Supposing the model state is composed of n_a modes of the atmospheric stream function ψ_a and temperature anomaly θ_a and n_o modes of the oceanic stream function ψ_o and temperature anomaly θ_o, respectively, the model state is given as

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

The total number of variables is $2 \times n_a + 2 \times n_o$. The key feature of MAOOAM is that we can change the resolution of the model by simply modifying the number of atmospheric and oceanic model variables.

In this study, we make use of two configurations of MAOOAM the same as Brajard et al. (2021): one with 56 variables ($n_a = 20$, $n_o = 8$, hereafter referred to as **M56**) and the other one with 36 variables ($n_a = 10$, $n_o = 8$, hereafter referred to as **M36**). Note that the configuration **M36** has the same resolution in the ocean component as the configuration **M56**, but 10 modes less in the atmosphere. These missing modes represent the high-order atmospheric modes and lead to the fact that **M36** does not resolve variability on small scales. Therefore, the model error in this study primarily comes from unresolved small-scale processes.

2.2 Ensemble Kalman Filter

The EnKF is a flow-dependent and multivariate DA method and has been implemented for climate prediction (Karspeck et al., 2013; Wang et al., 2019; Zhang et al., 2007). In the EnKF, the covariance is constructed from the dynamical ensemble and is more reliable than a static covariance (Sakov & Sandery, 2015). In addition, the ensemble-based covariance makes the updates satisfy the model dynamics and limits the assimilation shocks (Evensen, 2003).

In this study, we employ the DAPPER package (Raanes, 2018) to carry out the assimilation experiment. The DAPPER package is a toolbox for evaluating the performance of DA methods. The package provides experimental support and guidance for new developments in DA. We use the finite-size ensemble Kalman filter (EnKF-N) (Bocquet et al., 2015), which is the same method used by Brajard et al. (2021). One reason for choosing the EnKF-N algorithm is its numerical efficiency. This method can also automatically estimate the inflation factor, which can facilitate the assimilation experiment since it is a critical parameter to tune in ensemble data assimilation systems. We do not expect that using the traditional EnKF changes any of the conclusions of this paper. Therefore, in the following, we do not distinguish the EnKF-N from the traditional EnKF (hereafter the EnKF).

2.3 Artificial Neural Network Architecture

We suppose the dynamical model prediction is expressed as follows:

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

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 t_k to t_{k+1}. The model error at t_{k+1} is defined as follows:

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

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

We aim to use ANN to emulate the model error ε_{k+1}. Our ANN configuration is the same as in Brajard et al. (2021). The ANN architecture is composed of dense layers and the activation function is a linear rectification function (denoted "reLU"). Some additional parameters have been added, mainly to regularize the training: a batch norm layer at the input layer, which normalizes the training batch, and an L2-regularisation term on the parameters of the last layer. The parameters of ANN are optimized using the "RMSprop" optimizer over 300 epochs. For details, please refer to Brajard et al. (2021).

The error surrogate model can be expressed as follows:

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

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

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

2.4 Experimental settings

We present our experiments in Figure 1. The experiments are based on the two configurations of MAOOAM described in section 2.1. We define the model configuration with 56 variables (i.e., **M56**, section 2.1) as the true climate system and the model configuration with 36 variables (i.e., **M36**) as a dynamical prediction system. We carry out experiments (Figure 1) as follows:

- we integrate **M56** with a time step of approximately 1.6 minutes over 30726.5 years which is considered to as the spin-up period (De Cruz et al., 2016). We continue the simulation over 219 years which is defined as the "truth". We generate observations every 27 hours (i.e., every 10 time steps) by perturbing the "truth" using a Gaussian random noise with a standard deviation equal to 10% of the temporal standard deviation of the true state after subtracting the one-month running average ($\boldsymbol{\sigma}^{\mathrm{hf}}$).

- we perform a simulation with 50 ensemble members. 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 produce an analysis dataset with an ensemble size of 50.
- We produce two sets of ensemble predictions with 50 members: one with the dynamical model (i.e., **M36**) and the other with the hybrid model. The predictions start each second year from the year 125 to the year 185, last for 30 years, and have 50 ensemble members. Their initial conditions are taken from the analysis in the validation period (see Figure 1).

We split the analysis into two parts:

- Training part: The former 124.6 years of the dataset is used to train the parameters of the ANN, and apply the parameters to build the hybrid model.
- Validation part: The latter 94.6 years of the dataset is used to validate the ANN training and initialize prediction experiments (Figure 1).

Note that we utilize the same configurations of the model, DA, and ANN as Brajard et al. (2021). However, our experiments are different from that in Brajard et al. (2021) as follows:

- Brajard et al. (2021) performed an analysis experiment about 62 years. They used these data for both ANN training and validation. Here, we extended the simulation time to 219.2 years. And we divided the data into two separate parts: training and validation.
- Brajard et al. (2021) used the truth to initialize predictions. In our experiments, we use the analysis as initial conditions, which is more realistic because initial conditions are never perfectly known in reality.
- Brajard et al. (2021) performed predictions with one member by assessing one lead time only. We use the ensemble prediction with 50 members at several lead times.

2.5 Validation metrics

To test the prediction skill of the hybrid model, we adopt a metric commonly used in weather and ocean forecasting and climate prediction: the skill score (SS) (Murphy, 1988). The metric SS is based on the ensemble mean of the prediction and is defined as:

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

Here, $\mathrm{RMSE}_{\mathrm{prediction}}$ represents the Root Mean Square Error (RMSE) of the prediction (ensemble mean) against the truth, where the prediction is the result of the dynamical model or hybrid model. $\mathrm{RMSE}_{\mathrm{persistence}}$ represents the RMSE of the persistence prediction (in which the state at any lead time is the same as the initial conditions) against the truth. A positive SS indicates the prediction is better than the persistence and is skillful. A negative SS indicates the prediction is worse than the persistence and is not skillful. One advantage of the SS is that it is unitless. Thus, the SS is suitable for validation across different variables in the same panel (e.g., Figure 2).

For the significance test of the SS, we use a two-tailed Student's t-test to test the difference between the mean squared errors of the prediction and persistence. We use the bootstrap method to estimate the uncertainties of the SS. Since the SS is based on 30 prediction experiments, we randomly select (with replacement) 30 data from the 30 prediction experiments. Then we calculate the SS with these 30 sampled data. After repeating this procedure 10,000 times, we obtain a sample of 10,000 SS values and make use of their standard deviation as the uncertainties of SS.

3 Result

3.1 Prediction skill

Figure 2a shows the prediction skills of the dynamical model for both atmospheric temperature θ_a and stream function φ_a. We find that the variables in low-order atmospheric modes such as $\varphi_{a,2}$, $\varphi_{a,3}$, $\theta_{a,2}$ and $\theta_{a,3}$ have significant prediction skills until 14 days. While the temperature in high-order modes has significant skills within 8 days, the stream function in high-order modes has no prediction skill at all times. Figure 2b shows the prediction skills of the hybrid model for atmospheric variables. For temperature, the hybrid model is skillful for up to 18 days for all modes. For stream function, the hybrid model is skillful in predicting low-order atmospheric modes for up to 20 days and high-order modes for up to 14 days (exceptionally, $\varphi_{a,9}$ up to 20 days). Overall, the hybrid model is significantly more skillful than the dynamical model for atmospheric variables.

In the coupled model, the purpose of introducing ML to correct model errors is not only to improve the short-term atmospheric prediction skills (less than 14 days) of the model but also to improve the long-term oceanic prediction skills (over 5 years) of the model.

Figure 2c shows the prediction skills of the dynamical model for oceanic temperature and stream function. Since the ocean has lower variability than the atmosphere, the dynamical model has significant prediction skills for up to 30 years in oceanic temperature in most modes and oceanic stream function in some modes. In addition, the temperature is more predictable than the stream function. Figure 2d presents the prediction skills of the hybrid model. The hybrid model has significant prediction skills in both oceanic temperature and stream function in all modes for up to 30 years. It is worth noting that the hybrid model has higher SS than the dynamical model, in particular, for ocean temperature in the first and last modes and some oceanic stream functions in which the dynamical model has no prediction skill at all (e.g., $\varphi_{o,2}$ and $\varphi_{o,6}$).

Supporting information S1-S4 are examples of restoring variables in the physical space. The results also show that compared to the dynamical model, the hybrid model is closer to the truth in terms of spatial distribution and evolution. For long-term climate prediction, there are additional requirements for the hybrid model: the model must be able to can run for a long time and not diverge (Brenowitz et al., 2020; Rasp, 2020). In our case, there is no significant physical instability in the hybrid model during the predictions of 30 years. Overall, the hybrid model outperforms the dynamical model, which demonstrates the benefit of the data-driven error correction model built by the ANN.

3.2 Sensitive experiments

In the previous section, ANN is trained with the inputs from atmospheric and oceanic variables to correct both atmospheric and oceanic errors. In this section, we build two other hybrid models in which ANN is trained with the same input as the previous section but to correct either only atmospheric errors or only oceanic errors. The idea is to identify the error of which component is most important for predictions. We explore the prediction skills of three key variables of MAOOAM (De Cruz et al., 2016): $\varphi_{a,1}$, $\varphi_{o,2}$ and $\theta_{o,2}$.

Figure 3a shows the prediction skill of different hybrid models for the key atmospheric variable $\varphi_{a,1}$. Correcting both atmospheric and oceanic errors (the cyan line in Figure 3a) and correcting only atmospheric (the purple line in Figure 3a) have almost no significant difference. However, compared with the dynamical model result (the black dashed line in Figure 3a), correcting only the oceanic errors (the blue line in Figure 3a) does not improve the atmospheric prediction within 20 days.

Figure 3b and 3c show the prediction skill of different hybrid models for the two key oceanic variables $\varphi_{o,2}$ and $\theta_{o,2}$. Correcting both atmospheric and oceanic errors (cyan line) has the best prediction skill. Correcting only oceanic errors (blue line) can improve the prediction skill, but significantly less efficient than correcting both atmospheric and oceanic errors. For $\varphi_{o,2}$, when correcting only the errors in the ocean, there is a slight improvement in the first five lead years. But correcting atmospheric errors does not improve prediction skills in the first five years. It is mostly because of the physical unbalance between the atmosphere and the ocean and the fact that the ocean needs some time to synchronize with the error-corrected atmosphere. For $\theta_{o,2}$, correcting only oceanic errors (the blue line) and only atmosphere errors (the red line) show high SS in the first 15 years.

4 Conclusions and Discussions

In this study, we applied a method to online correct the model error of a simplified ocean-atmosphere coupled model (MAOOAM). The ML is introduced to learn the model error between the analysis performed by DA and the hindcast simulation thus building a statistical-dynamical hybrid model. The hybrid model is able to make reasonable prediction skills using both the atmospheric and oceanic model states as input. Besides, we find if we only focus on improving short-term prediction skills of atmospheric variables, only correcting the atmospheric error can obtain a similar prediction skill by correcting both atmospheric and oceanic errors. But good prediction skills for ocean variables require correction for both atmospheric and oceanic model errors.

This study is to be seen as a proof of concept, in which we have shown that in principle it is possible to let ANN learn the model error and thus improve the prediction skills of the coupled model. Ideally, one would apply the ML corrections to the same model that is used to generate the analysis. This also effectively solves the problem of how to correct the model error when the observation is insufficient and cannot be directly used for ML training. In an operational weather forecasting context, it would be possible to adapt this method to learn model errors from a fully-fledged DA system which would ensure consistency between the models.

Besides, a realistic model is more complex than MAOOAM and the correcting frequency in a realistic model is lower. The next natural step for future studies would apply this method to the realistic model and explore the prediction skills.

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. Figures were made with Matlab version 2018a. MAOOAM (Demaeyer et al., 2020) is available at https://github.com/Climdyn/qgs. Dapper version 0.9.6 (Raanes, 2018) is available at https://github.com/nansencenter/DAPPER/tree/v1.3.0.

Acknowledgments

This was funded by the National Key *R&D* Program of China (2022YFE0106400), 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 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.

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375.

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.

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.

Demaeyer, J., Cruz, L. D., & Vannitsem, S. (2020). qgs: A flexible python framework of reduced-order multiscale climate models. *Journal of Open Source Software*, *5*(56), 2597. Retrieved from `https://doi.org/10.21105/joss.02597` doi: 10.21105/joss.02597

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.

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108.

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.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, *116*(12), 2417–2424.

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.

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

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1. 0). *Geoscientific Model Development*, *13*(5), 2185–2196.

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

Salcedo-Sanz, S., Ghamisi, P., Piles, M., Werner, M., Cuadra, L., Moreno-Martínez, A., ... Camps-Valls, G. (2020). Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources. *Information Fusion*, *63*, 256–272.

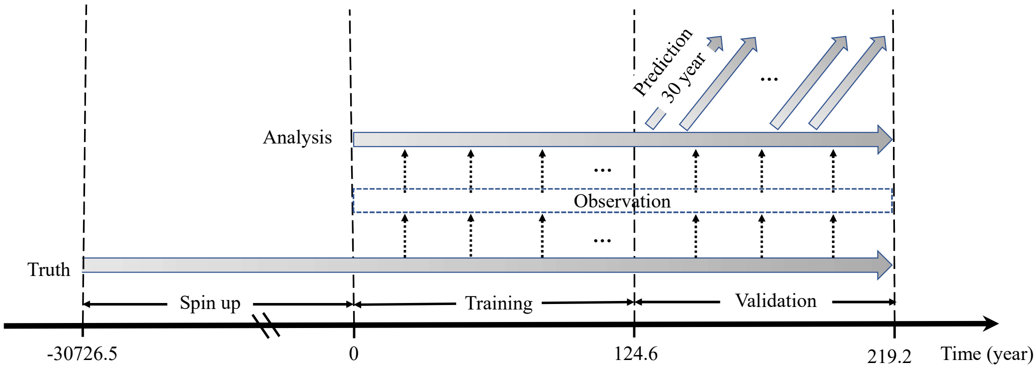

Figure 1. Schematic of experiments listed in section 2.4.

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.

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.

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.

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.

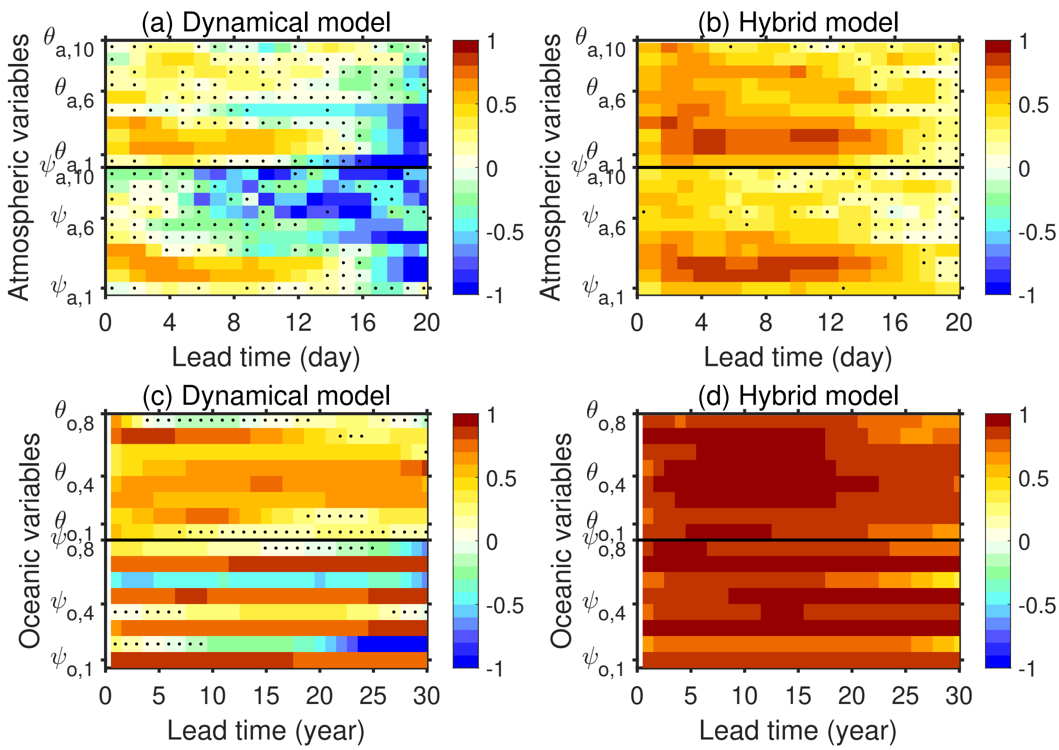

Figure 2. SS as a function of the prediction lead time for variables in the hybrid model or the dynamical model. (a) The SS of the dynamical model for atmospheric variables, (b) the SS of the hybrid model for atmospheric variables, (c) The SS of the dynamical model for oceanic variables, and (d) the SS of the hybrid model for oceanic variables. The black dot indicates the SS not exceeds the 95 significance test.

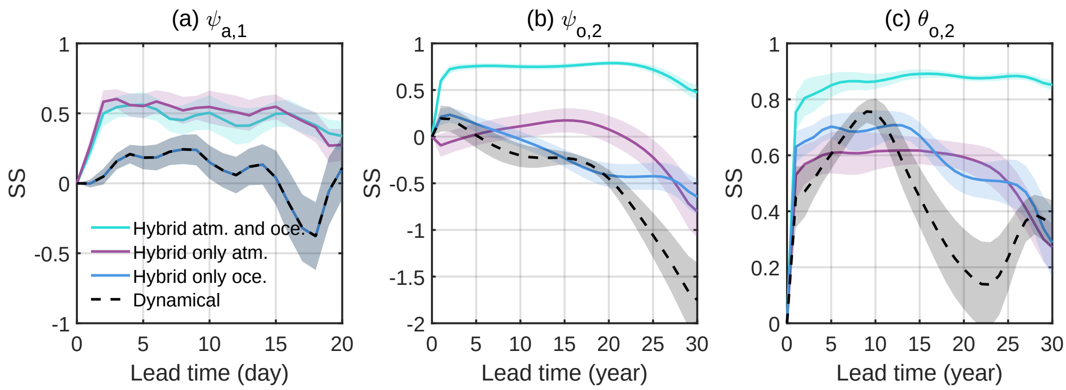

Figure 3. SS for three key variables (a) $\psi_{a,1}$, (b) $\psi_{o,2}$ and (c) $\theta_{o,2}$ as a function of lead time (20 days for the atmospheric variable and 30 years for the ocean variables). Shading shows one standard deviation calculated by the bootstrap method described in section 2.5. The cyan line is the SS of the hybrid model built by correcting both atmospheric and oceanic model errors, the purple line is the SS of the hybrid model built by only correcting atmospheric model errors, the blue line is the SS of the hybrid model built by only correcting oceanic model errors and the dash black line is the SS of the dynamical model.