Responses to Reviewers' Comments for Manuscript egusphere-2025-212

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

Addressed Comments for Publication to

Nonlinear Processes in Geophysics

by

Zikang He, Julien Brajard, Yiguo Wang, Xidong Wang, Zheqi Shen

Authors' Response to Reviewer #1

Comment 1

By using a reduced-order coupled atmosphere-ocean model within a twin experiment framework, the authors demonstrate the capability of machine learning in correcting errors in various components of the coupled model on different time scales so that enhancing the prediction skill of the model. I found the results interesting and publishable and recommend a minor revision that addresses my comments below.

Response: We thank the reviewer for providing insightful comments that have helped to significantly improve the manuscript. We carefully addressed each concern and revised the manuscript. Below, we provide our detailed point-by-point responses to the reviewer's comments. To enhance the legibility of this response letter, all the reviewer's comments are typeset in blue boxes.

Specific Comments:

Comment 2

Line 22: "estimate best the state" should be "estimate the best state".

Response: We thank the reviewer for this comment. We have revised it to "estimate the best state" (L23 in the manuscript).

Comment 3

Line 127: Section 2.4. Experimental settings, "experimental settings" sounds like the setting itself is experimental, "experiment settings" sounds more like setting up an experiment.

Response: We thank the reviewer for this comment. We have revised it to "Experiment Settings" (L148 in the manuscript).

Responses to Reviewers' Comments for Manuscript egusphere-2025-212

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

Addressed Comments for Publication to

Nonlinear Processes in Geophysics

by

Zikang He, Julien Brajard, Yiguo Wang, Xidong Wang, Zheqi Shen

Authors' Response to Reviewer #2

Comment 1

In this manuscript the authors use hybrid modelling to correct model error in a low-order coupled ocean-atmosphere model, namely MAOOAM. Model error is introduced by using a spectral truncation of a reference version of the model. Then, a neural network is trained to correct the 27h forecast errors by learning the analysis increments obtained from an EnKF analysis. Finally, the hybrid model is evaluated in ensemble forecast experiments.
Overall, the manuscript reads well, but I have the feeling that some key informations / explanations about the experiments are missing (see general points).
Nevertheless I think that after substantial revision it could make a valuable contribution to the literature.

Response: We thank the reviewer for providing insightful comments that have helped to significantly improve the manuscript. We carefully addressed each concern and revised the manuscript. Below, we provide our detailed point-by-point responses to the reviewer's comments. To enhance the legibility of this response letter, all the reviewer's comments are typeset in blue boxes. Rephrased or added sentences in the revised version are indicated in a gray box.

General comments

Comment 2

1) At first glance, I had difficulties to grasp what are the objectives of this work, and to what extent it differs from previous work on hybrid modelling (in particular Brajard et al 2021). Therefore, I think that it would make sense to clarify a bit the objectives of this work (in the introduction) and in particular what it means for you to built a model error correction for climate prediction and how it differs from eg NWP or dynamical systems in general.

Response: We thank the reviewer for this comment. For clarity, we have revised the introduction as follows (L65–L72 in the manuscript) to better highlight the objectives of the study:

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

We have also explicitly emphasized how our experimental setup differs from that of Brajard et al. (2021), particularly in the use of imperfect initial conditions and long-term climate prediction settings (L172–L193 in the manuscript):

> It is worth noting that since we employ the same ANN configurations as outlined in Brajard et al. (2021), the ANN parameters in this study are trained only once, without any modifications throughout the training process by using a separate validation set. We examined the loss curves (not shown in this study) to assess

the training behavior. The loss curves provided evidence that the network was continuing to learn without signs of overfitting throughout the training process.
Brajard et al. (2021) focused on developing the hybrid model methodology; our study aims to explore the evolution of prediction skill as a function of lead time. We assess the prediction skill over a wider range of lead times, specifically up to 50 days for atmospheric variables and up to 60 years for oceanic variables. By examining the skill at various lead times, we can gain insights into the temporal evolution and long-term performance of the hybrid model, providing a more comprehensive understanding of its capabilities and limitations. To do so, our experimental setup is different from that of Brajard et al. (2021) in the following ways:

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

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

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

Comment 3

2) It is absolutely necessary to give typical time scales of the MAOOAM model for each component (atmosphere and ocean), for example by providing the doubling time of errors. Without this information, it is really hard to get an idea of the time evolution in the model and for example to know whether the 50 days and 60 years lead time used in some experiments are long or not.

Response:

We thank the reviewer for this helpful suggestion. As suggested, we have added the analysis on the doubling time of errors as follows (L240-L249 in the manuscript):

The distinction between short-term (daily) and long-term (yearly-averaged) prediction scales in this study is based on the fundamentally different error growth characteristics of atmospheric and oceanic variables. As illustrated in Fig. R1, atmospheric variables exhibit rapid error amplification, with a doubling error time of approximately one day and saturation occurring within about ten days. In contrast, oceanic variables demonstrate much slower error growth, with errors roughly doubling over the first year and continuing to grow gradually over the subsequent decade.

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

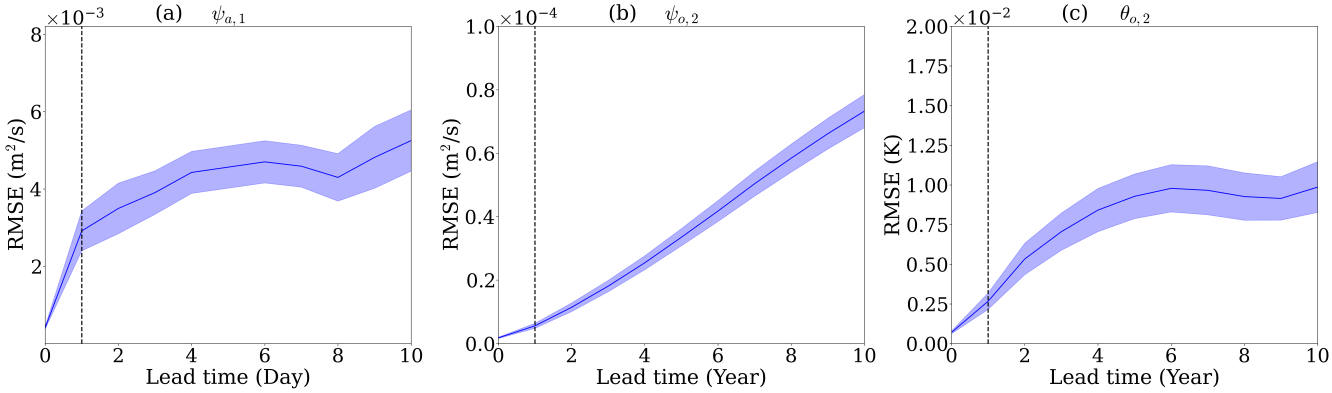

Figure R1. RMSE of dynamical prediction as a function of lead time for three key variables in spectral space: (a) $\psi_{a,1}$, (b) $\psi_{o,2}$, and (c) $\theta_{o,2}$. The atmospheric variable ($\psi_{a,1}$) is evaluated based on instantaneous outputs sampled every 27 hours, while the oceanic variables ($\psi_{o,2}$ and $\theta_{o,2}$) are evaluated using yearly averages. Shading indicates one standard deviation, representing the uncertainty of prediction skill, estimated using the bootstrap method. The vertical dashed lines represent the time of doubling error.

Comment 4

3) I find the use of the significance test really difficult to understand. In particular, I don't understand what it means for a correlation or for a skill score "to be significant"? For example, what would be the difference between a significant 50% correlation and a non-significant 50% correlation? Perhaps it would be clearer if you explicitly indicated what are the null and tested hypotheses in your Student's t-test. More generally, I have the feeling that you would get a similar information content by computing a confidence interval over the 30 test runs, with the added benefit that it would be much easier to understand and hence to explain.

Response: We agree that the use of statistical significance testing warrants clarification. We have revised the manuscript as follows (L211-L221 in the manuscript):

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

as normally distributed:

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

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

Comment 5

4) The caption of figure 3 mentions "The atmospheric variables are calculated based on daily data, while the oceanic variables are based on annual average data". Does this mean that the truth and the prediction are averaged (on a daily or annual basis) before computing the RMSE or correlation? To me, this is a really important point, because it means that you evaluate the models in their ability to reproduce daily or annual means, and not their ability to predict trajectories. This is fine, especially in a context of "climate predictions" but should absolutely be discussed in the main text. In my opinion, this raises additional questions, like for example why only the mean, why not higher-order moments or more complex statistical properties? Also, it is not entirely clear to me how fast is the evolution of daily / annual means in this model, and hence whether the averaging window (one day or on year) makes sense or not.

Response: In this study, the instantaneous model outputs are restored every 27 hours. Therefore, for atmosphere evaluations (e.g., correlation and RMSE-SS), we use these instantaneous outputs. For ocean evaluation, we compute the metrics based on the annual mean values, rather than using a snapshot. We have revised the manuscript as follows (L228-L237 in the manuscript):

In climate prediction, time-mean quantities such as monthly (Wang et al., 2019) or annual averages (Boer et al., 2016; Bethke et al., 2021) are often used because time averaging reduces the impact of chaotic weather variability, making the underlying climate signals more apparent. They also better meet the practical needs of sectors such as agriculture and energy, where planning is often based on mean conditions. In contrast, predicting higher-order statistics accurately remains challenging due to model limitations and computational costs, particularly when high-resolution Earth system models are required. Nevertheless, there is growing interest in representing complex statistical properties to improve the prediction of extreme events and support climate risk assessments.
To evaluate prediction skill across different time scales, we apply two complementary strategies. For short-term predictions, instantaneous outputs sampled every 27 hours are used to represent daily variations. For long-term predictions, model outputs are averaged annually to assess the ability to capture low-frequency variability.

Comment 6

5) The scientific question raised in section 3.2 is interesting, but I am not entirely sure that the experimental setup chosen in this manuscript is entirely relevant to answer that question, because there is by construction no model error on the oceanic component (which means that the only source of error in the ocean comes from the interaction with the atmosphere). Could you discuss this point?

Response: We thank the reviewer for the insightful comment. We agree with the reviewer that the current experimental setup can not entirely address the importance of atmospheric or oceanic error correction in climate prediction. Ideally, we should also use an imperfect ocean component. For clarity, we have revised the manuscript as follows (L290-L294 and L336-L344 in the manuscript):

> 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 gain some insight into the relative importance of atmospheric and oceanic error correction for the overall performance of the climate prediction on different time scales.

> This study also examined the respective impacts of correcting atmospheric and oceanic model errors on prediction skill. Our results indicate that short-term atmospheric predictions are primarily influenced by atmospheric model errors, while correcting oceanic errors alone has a limited effect (e.g., Balmaseda and Anderson, 2009). For long-term ocean prediction, correcting atmospheric errors is essential due to their role in surface forcing, while correcting oceanic errors plays a more critical role in predicting ocean temperature. It is worth noting that in our experiment setup, the ocean component is perfect, and its prediction errors primarily come from the errors in the atmospheric component. However, correcting the ocean model errors can influence the atmosphere through the coupling between the ocean and the atmosphere. Although the experimental setup is not ideal, our results still provide some insights into the relative importance of oceanic error correction for the prediction on different time scales.

Specific and technical comments

Comment 7

- Introduction: I have the impression that you move back and forth between dynamical models in general and climate models in particular, which makes the text sometimes a bit harder to follow.

Response: We thank the reviewer for this helpful comment. In the revised manuscript, in particular in the abstract and introduction sections, we have carefully reviewed and unified the terminology by consistently using "dynamical climate model" instead of "climate model" or "dynamical model" to improve the clarity and coherence of the text.

Comment 8

- L 22 "estimate best the state of the climate system" there is a typo in this sentence. Furthermore, I would rather use "initial condition" here than "state" (to be consistent with the start of the sentence).

Response: We have revised the sentences in the manuscript as follows (L20-L23 in the manuscript):

> To reduce the uncertainties of initial conditions, climate prediction centers (Balmaseda and Anderson, 2009; Doblas-Reyes et al., 2013) have been evolving towards the use of data assimilation (DA, Carrassi et al., 2018) which combines observations with the dynamical climate models to estimate the best initial conditions of the climate prediction (Penny and Hamill, 2017).

Comment 9

- L 39-40 "In these works, the hybrid model is tested in an idealized setting in which initial conditions are perfectly known" I guess that here you are referring to the fact that the forecast skill of the hybrid models are evaluated using a dataset with perfect initial conditions, but the current formulation is not entirely clear to me.

Response: Since the text has been revised, this comment is no longer relevant, but we thank the reviewer for the comment.

Comment 10

- L 41-43 "To our knowledge, the performance of hybrid models under imperfect initial conditions—particularly when using an ensemble of forecasts—has not been thoroughly assessed." Actually, starting the forecast from an analysis and not necessarily perfect initial conditions has already been done, eg in Farchi et al, 2023 (doi 10.1029/2022MS003474).

Response: We thank the reviewer for the comment. For clarity, we have revised the manuscript as follows (L41-L43 in the manuscript):

> While Brajard et al. (2021) conducted prediction experiments using perfect initial conditions, more recent studies such as Farchi et al. (2023) examined the performance of hybrid models initialized with imperfect conditions, using a two-layer quasi-geostrophic (QG) model.

Comment 11

- L 47-48: "Moreover, observation for training, validation, and testing is relatively limited." I don't fully agree with this. For specific components of the Earth system (eg the atmosphere), there are a lot of observations available.

Response: We agree with the reviewer that atmospheric observations are far more abundant than those for other components of the climate system, such as the ocean. However, we intended to highlight a different challenge—namely, the limited availability of long-term observational records, particularly before the satellite era. This scarcity poses a significant constraint for studies focused on decadal or long-term climate prediction, where extended and consistent datasets are crucial for ML-based model development and evaluation. For clarity, we have revised the manuscript as follows (L61-L64 in the manuscript):

However, the potential benefits of ML-based model error correction for climate prediction across different time scales remain largely unexplored. This is primarily due to the sparsity of long-term observational records (such as those spanning the 20th century) in both time and space, which presents significant challenges for developing effective ML-based error correction models for climate prediction applications.

Comment 12

- L 71-75 Why is the relationship between resolution and number of modes different in the atmosphere and in the ocean?

Response: We thank the reviewer for this question. For clarity, we have added the following statements into the manuscript (L93-L96 in the manuscript):

The total number of variables in the model state is $2n_a + 2n_o$. Note that n_a is typically larger than n_o, reflecting the distinct characteristics of the two components in MAOOAM. The atmosphere exhibits faster dynamics and smaller-scale variability, necessitating a greater number of modes to adequately capture its behavior. In contrast, the ocean evolves more slowly and is dominated by larger-scale processes, which can be effectively represented using fewer modes (De Cruz et al., 2016).

Comment 13

- L 77-79 "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." Actually, any model with a spatial extent possesses this feature, right?

Response: We agree that the ability to modify the number of model variables by adjusting spatial resolution is not unique to MAOOAM. To address this comment, we have revised the sentence to better reflect the specific advantage of MAOOAM. We have revised the manuscript as follows (L98-L100 in the manuscript):

Like many other models formulated in spectral space, MAOOAM offers flexibility in adjusting the number of atmospheric and oceanic variables by simply modifying the model resolution in spectral space.

Comment 14

- Figure 1 is difficult to interpret. Beyond the fact that the attractors look different, I don't get much out of it. I would advise first to use the same axes extent for both panels, to use label fonts consistent with the main text, and to use colour or transparency to show the density. Beyond these advises, is the 3D aspect really fundamental here? Otherwise, I would suggest to show a 2D figure (potentially with more panels if needed), which would be easier to interpret.

Response: We thank the reviewer for this comment. For clarity, we have revised the figure as follows:

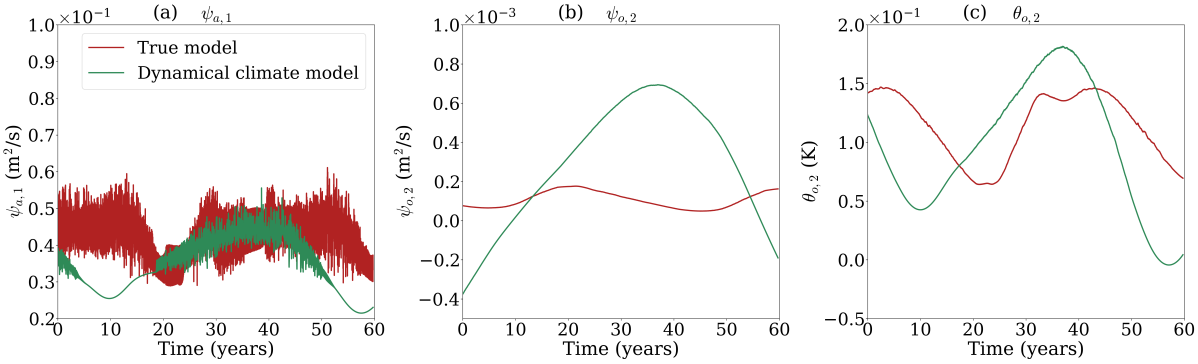

Figure R2. Time series of the true model (red lines) and the dynamical climate model (green lines) for three key variables: (a) $\psi_{a,1}$, (b) $\psi_{o,2}$, and (c) $\theta_{o,2}$.

Comment 15

- L 86 "showing they evolve differently" If I am not mistaken there is a typo in this part of the sentence.

Response: As suggested, we have revised the manuscript as follows (L106-L108 in the manuscript):

> Figure 1 displays time series of three key variables in the true model M56 and the dynamical climate model M36 in spectral space, illustrating their different evolution patterns (De Cruz et al., 2016).

Comment 16

- L 89 "specifically 10 mode less" -> "specifically 10 modes less"

Response: We have corrected "10 mode less" to "10 modes fewer" in the revised manuscript (L110 in the paper).

Comment 17

- L 90 "The atmospheric error could then propagate" by using "could" instead of something more affirmative, you imply that in certain cases the error can be limited to the atmospheric component?

Response: In our model configuration, the components are fully coupled, and atmospheric errors inevitably affect the ocean. We have revised the manuscript as follows (L111-L114 in the manuscript):

> The atmospheric error in the y-direction propagates to the atmosphere in the x-direction and the ocean component through the coupling terms in the equations.

Comment 18

- L 98-99 "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." You should reformulate this sentence, because as is it sounds like section 2.4 and figure 2 will be about the use of DAPPER.

Response: We thank the reviewer for these comments. We have revised the manuscript as follows (L120-L121 in the manuscript):

> All experiments in this study are conducted using the DAPPER package (Raanes, 2018). The overall experimental setup is described in section 2.4 and depicted in Fig. 2.

Comment 19

- L 100-101 "This method reducing the amount of experimentation required in tuning the EnKF DA system, thereby enhancing the performance of the assimilation experiments" I would reformulate this sentence, as one could understand that reducing the number of experiments enhances performance, which is obviously not the case.

Response: We agree that the original sentence was misleading as it implied a causal relationship between reducing the number of experiments and improving performance. Our intended meaning was that the method adaptively adjusts the inflation factor, which minimizes the need for manual tuning. We have revised the manuscript as follows (L122-L123 in the manuscript):

> This method adaptively adjusts the inflation factor, thereby reducing the need for extensive manual tuning and enhancing the performance of the assimilation experiments.

Comment 20

- L 111-112 "the training of ANN is using the analysis increments produced by the EnKF (Gregory et al., 2024)" I would suggest to cite Farchi et al 2021 (doi 10.1002/qj.4116) and Brajard et al 2021 (which you already cited elsewhere) before Gregory et al 2024 here.

Response: We agree that it is more appropriate to cite some earlier key works before citing Gregory et al. (2024). We have revised the manuscript as follows (L132-L134 in the manuscript):

> We aim to use ANN to emulate the model error ε_{k+1}. Since the truth is not known in practice, the training of ANN uses the analysis increments produced by the EnKF (Brajard et al., 2021; Farchi et al., 2021; Gregory et al., 2024).

Comment 21

- L 126 Is the NN correction applied in spectral space or in grid-point space?

Response: The NN correction is applied in the spectral space. For clarity, we have revised the manuscript as follows (L166-L167 in the manuscript):

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

Comment 22

At Line 135, please replace "true state σ^{hf}" with "true state $\mathbf{x}^{\mathrm{t}}$" to ensure consistency with the notation introduced at Line 110.

Response: Thanks to the reviewer for the comment. We have revised the manuscript as follows (L156-L157 in the manuscript):

> The standard deviation ($\boldsymbol{\sigma}^{\mathrm{hf}}$) of the noise is set to 10% of the temporal standard deviation of the true state ($\mathbf{x}^{\mathbf{t}}$) after subtracting the one-month running average.

Comment 23

- L 132-136 You forgot to specify what is the observation operator (H=I I imagine) and whether it is applied in spectral space or in grid point space?

Response: We thank the reviewer for pointing this out. Indeed, the observation operator is the identity matrix and is applied in spectral space. To clarify this aspect, we have revised the manuscript as follows (L157-L158 in the manuscript):

> Observations are generated every 27 hours in spectral space, while the observation operator H is the identity operator ($H = I$) and is also applied in spectral space.

Comment 24

- L 137 "and generate a reanalysis" Why a reanalysis here? Why not simply an analysis?

Response: Sorry for the confusion. We have replaced "reanalysis" with "analysis" (L159).

Comment 25

- L 150 "without incorporating validation data to adjust the ANN model during training" If you don't have any validation set, how can you be sure that the training process has converged?

Response: We adopted the same ANN configuration as described in Brajard et al. (2021), although their study addressed a different objective within the MAOOAM framework. In our implementation, the ANN was trained once for 300 epochs without incorporating a separate validation set to adjust the model during training. Instead, we evaluated the training process by monitoring the loss curves for both the training and test datasets. These curves showed that the network continued to learn throughout the training, with no indication of overfitting (Figure R3). While no explicit validation set was used for early stopping or hyperparameter tuning, we followed the original setup from Brajard et al. (2021) to ensure comparability and relied on post-training diagnostics to confirm convergence. We have revised the manuscript as follows (L172-L175 in the manuscript):

> It is worth noting that since we employ the same ANN configurations as outlined in Brajard et al. (2021), the ANN parameters in this study are trained only once, without any modifications throughout the training process by using a separate validation set. We examined the loss curves (not shown in this study) to assess the training behavior. The loss curves provided evidence that the network was continuing to learn without signs of overfitting throughout the training process.

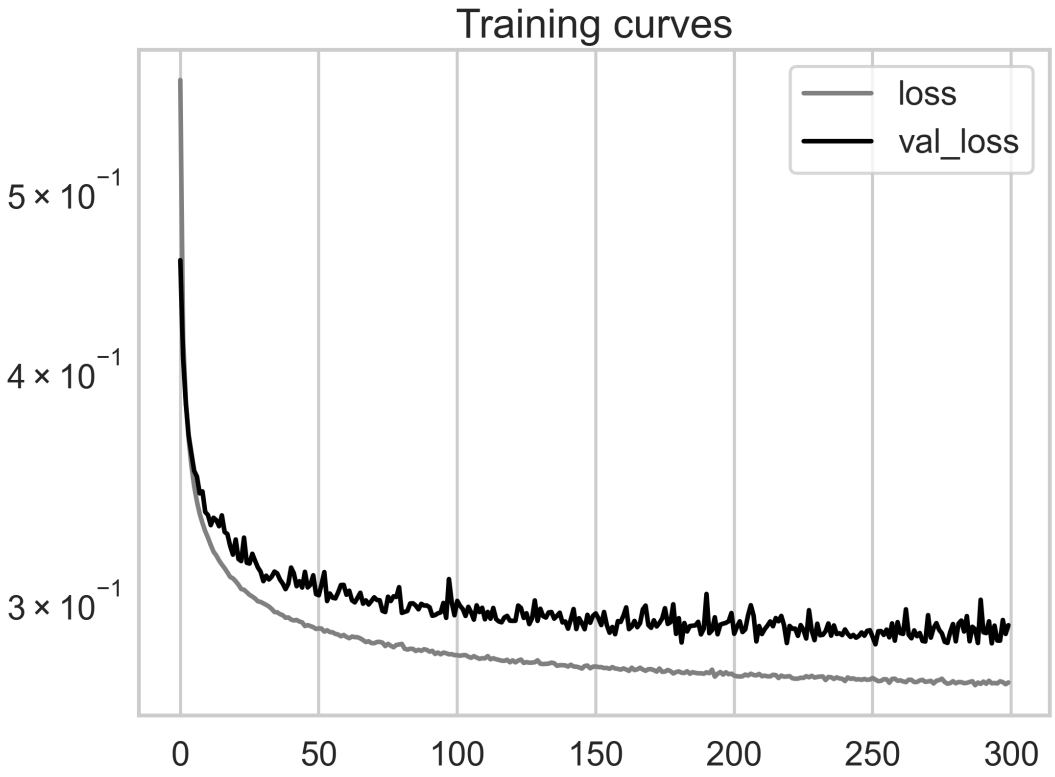

Figure R3. Training curves. The validation loss is computed by using the test dataset.

Comment 26

- Section 2.5: Usually, in the RMSE the "mean" is computed over the different variables that constitute on state, in such a way that there is only one RMSE for a prediction. However, as far as I understand, in your case the mean in the RMSE is computed over the 30 test runs, right? In such a way that you get one RMSE value per variable. This should be explicitly mentioned.

Response: Indeed, in our analysis, the RMSE is computed separately for each variable, and the averaging is performed over the 30 predictions rather than over all state variables. This approach allows us to assess the prediction skill for each variable of the coupled system individually. We have explicitly clarified this point in section 2.5 of the revised manuscript (L195-L210 in the manuscript).

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

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

where x represents the prediction (ensemble mean) and y represents the truth. N is the total number of prediction experiments and is equal to 30 (section 2.4).
The RMSE is calculated as follows:

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

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

$$\text{RMSE-SS} = 1 - \frac{\text{RMSE}_{\text{prediction}}}{\text{RMSE}_{\text{persistence}}}, \tag{3}$$

where $\text{RMSE}_{\text{prediction}}$ represents the RMSE between the prediction (ensemble mean) and the truth and $\text{RMSE}_{\text{persistence}}$ represents the RMSE between a persistence prediction (where the state remains the same as the initial conditions) and the truth. A positive RMSE-SS indicates that the prediction outperforms persistence and demonstrates skill. On the other hand, a negative RMSE-SS indicates that the prediction performs worse than the persistence and lacks skill. By utilizing the correlation and RMSE-SS, we can assess and compare the skill of the predictions generated by the dynamical climate model and the hybrid model across different variables within the same panel, as shown in Fig. 4.

Comment 27

- L 199: phi should be psi, right?

Response: The reviewer is right. We have modified ϕ as ψ (L270).

Comment 28

- Figures 3 and 4 (and the other ones as well): I am a surprised by the "noisiness" of the results, because with low-order models the scores usually look much smoother. Could this be coming from the fact that the test set contains only 30 runs? Would it be affordable to increase this number?

Response: We thank the reviewer for the suggestion. To evaluate whether the noisiness in the results stems from the small number of test samples, we increased the number of prediction experiments from 30 (starting every 2 years) to 60 (starting every year) over the 60-year test period. As shown in Fig. R4, the results remain mostly unchanged, indicating that the noisiness is not due to the sample size. For simplification, we remain with 30 prediction experiments. Nevertheless, we do not fully understand the sources of the noisiness, but speculate that it is due to high-frequency variability in the atmosphere.

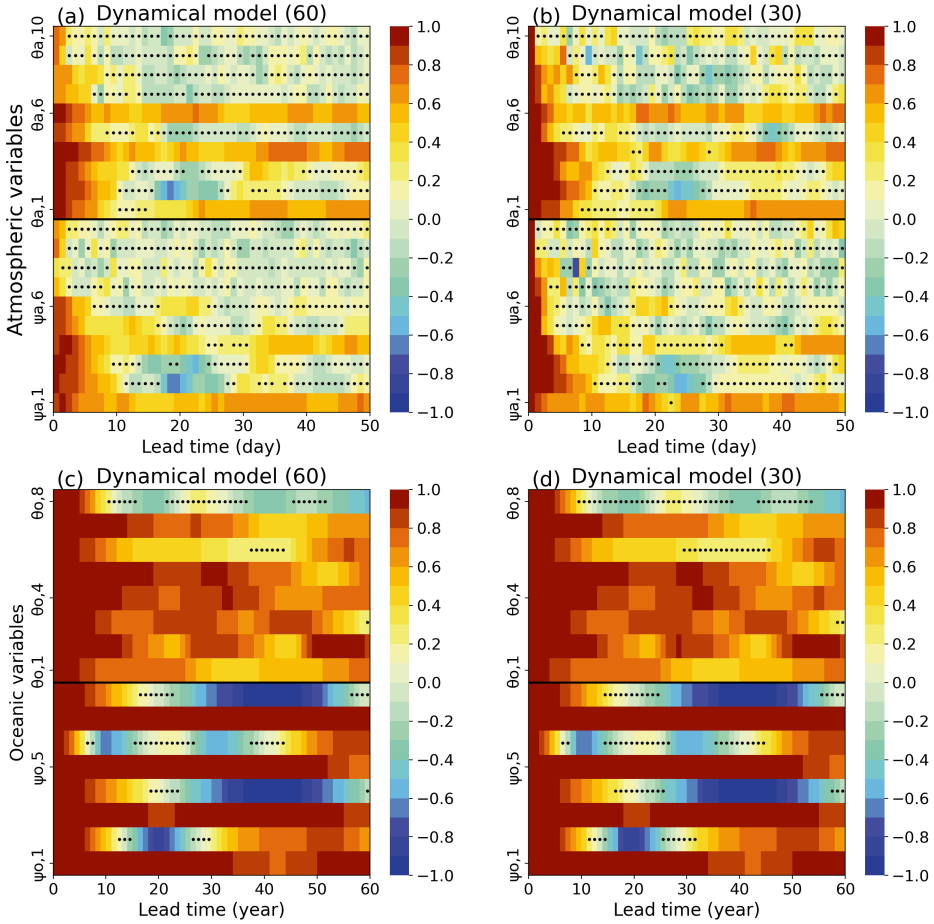

Figure R4. Correlation as a function of prediction lead time for different variables. Panels (a) and (c) are based on 60 prediction runs; panels (b) and (d) are based on 30 prediction runs. Atmospheric variables are evaluated using daily data, while oceanic variables are based on annual averages. Black dots indicate correlations that do not pass the 95% significance level.

Comment 29

- Figures 9 and 10: are the correlation and RMSE skill score computed again on daily / annual averages or on instantaneous values?

Response: We have revised the manuscript structure, and the original Figures 9 and 10 now correspond to Figure 7 in the updated version. In our experiments, the model restores instantaneous outputs every 27 hours of the model time. For atmospheric variables, the correlation and RMSE-SS presented in Figure 7 are computed directly from the 27-hour instantaneous outputs. For oceanic variables, annual means are first computed from the 27-hour instantaneous outputs, and the correlation and RMSE-SS are then calculated based on these annual means. We have added the relevant description in the revised manuscript as follows (L228-L234 in the manuscript):

> In climate prediction, time-mean quantities such as monthly (Wang et al., 2019) or annual averages (Boer et al., 2016; Bethke et al., 2021) are often used because time averaging reduces the impact of chaotic weather variability, making the underlying climate signals more apparent. They also better meet the practical needs of sectors such as agriculture and energy, where planning is often based on mean conditions. In contrast, predicting higher-order statistics accurately remains challenging due to model limitations and computational costs, particularly when high-resolution Earth system models are required. Nevertheless, there is growing interest in representing complex statistical properties to improve the prediction of extreme events and support climate risk assessments.
> To evaluate prediction skill across different time scales, we apply two complementary strategies. For short-term predictions, instantaneous outputs sampled every 27 hours are used to represent daily variations. For long-term predictions, model outputs are averaged annually to assess the ability to capture low-frequency variability.

Comment 30

- There is in section 3 a lot of (large) figures, which makes the text sometimes difficult to read (eg when referring to a figure which is located... in the following pages). I know that this is a draft and the one-column draft mode isn't really helping, but when moving to the edition process, I would recommend to be really careful to mitigate as much as possible the inconvenience.

Response: We thank the reviewer for this practical suggestion and fully agree that the current one-column draft layout makes it more difficult to follow the figures alongside the text. In response, we have revised the manuscript by merging similar figures to improve clarity. In the proof-reading version, we will carefully reorganize the layout and consult with the editorial office to ensure that figures are positioned as close as possible to the corresponding text, thereby enhancing readability and minimizing inconvenience.

References

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

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

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.

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.

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.

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, 2013.

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

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

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.

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.

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.