Shared Reply to All Reviewers

Dear reviewers,

Thank you for your constructive comments and review of our manuscript. Your feedback has helped us refine the content. Structurally, we have made several changes listed below:

1. Since SST initialization is involved when comparing GEPS5 and GEPS6, we change the title to “Persistent SST Anomaly vs Dynamical Ocean Model in Winter Weather Forecasts: Global Ensemble Predictions System Versions 5 and 6 over the North Pacific and North Atlantic”
2. We added more information about the SST initialization. Specifically, GEPS6 introduces an SST bias variance of $0.05\ K^2$ to the Kuroshio Current Extension region, and $0.5\ K^2$ to the Gulf Stream. We have also added text to provide potential reasons for such biases, including model resolution and air-sea interactions.
3. For better flow of the article, we have swapped Section 3.3 with Section 3.4, and Figure 4 with Figure 6.
4. We have added panels showing analysis of latent heat flux in new Figures 4d-f.
5. We have added global and hemispheric bias variance analyses of SST in Supplementary Figure 2 to provide information about the magnitude of initial SST bias variance that GEPS6 introduces on top of GEPS5.

We have addressed your comments (in red font) in this document, point by point. We thank you again for your time to improve this work.

Best,

Tien-Yiao

Reply to Reviewer 1's Comments

Reviewer #1

GENERAL COMMENTS

This paper compares ensemble forecasts of the GEPS6, which is dynamically coupled to NEMO, with hindcasts of GEPS5, which uses persisted SST anomalies. The authors find changes in the mean and variance between the two forecasts, both in the ocean and the atmosphere. The changes are discussed in terms of the effect of air-sea coupling.

While there is some interesting analysis here, I think the paper as it currently stands has several major problems. Some of this relates to the framing of the paper (the title makes it out to be about air-sea coupling but is actually about the broader effect of using a dynamical ocean model), and some of it relates to a potentially serious confounding effect (the initialisation of GEPS6 hindcasts is apparently very different from GEPS5 hindcasts). There are also many missing references to past literature.

I flesh out these and other issues in my comments below. Major revisions will be needed to address them. I look forward to reading a revised version.

Best wishes,

Kristian Strommen

MAJOR COMMENTS

1. The paper is framed in terms of the effect of air-sea coupling, but the comparison of the two forecasts is much better thought of as examining the impact of using a dynamically coupled ocean model. The distinction is important: the difference is not just that there is an exchange of information now between ocean and atmosphere, but that the dynamic ocean introduces its own unique SST biases (which it will have even when run without an atmosphere, since it's not a perfect model). Some of the changes documented in the paper seem to be about changes to the biases and not really about the two-way coupling. For example, the change in the Gulf Stream is consistent with the fact that NEMO at ¼ degree resolution does not simulate a Gulf Stream that separates from the continent correctly. This does not have anything to do with coupling (it happens also in ocean-only simulations), but is related to model resolution and bathymetry.

Studies that aim to really isolate coupling often deal with this by looking at things like lead-lag correlations between SSTs and wind-stress or fluxes, since correlations ignore magnitude and thus are insensitive (at least a priori) to model biases. For an example, see e.g. https://doi.org/10.1002/2016GL070559.

I think unless you want to almost completely redo the paper to follow similar methods, you need to reframe the paper to be much more specifically about the impact of using a coupled dynamical ocean model in your forecast. However, at this point it's clear the results depend sensitively on the exact model, since this determines the model biases. Thus, I think the authors should rephrase everything to be very specifically about the comparison between GEPS6 and GEPS5. This includes mentioning GEPS somewhere in the title. Air-sea coupling should not be mentioned in the title unless considerable additional analysis along the lines of the Roberts et al. paper (or similar) is added.

Reply: We thank the reviewer for the input. We concur that we should be careful with such details. Due to complications in the SST initialization and differing start dates, we decided to reframe the paper to focus on the impact of using a dynamic ocean model, as explained in the shared reply above.

2. Following on from the above, the relationship between this paper and the tech report of Lin et al. (2019) is highly unclear. It seems Lin et al. already look at the impact of forecast skill, but this is only mentioned all of a sudden in the middle of the discussion of your results. This surprised me given that Lin is a co-author here as well! You need to mention Lin et al. (2019) in the Introduction, and clearly discuss what their results are and how your analysis and results differ or complement theirs.

Reply: We apologize for not introducing the Lin et al. (2019, hereafter Lin19) findings earlier in the manuscript. Lin19 found that, GEPS6 forecasts better winter Arctic sea ice in the Pacific and Eurasian sectors, surface air temperature, tropical SST, and MJO activity. However, they provided less evaluation of the North Pacific and North Atlantic, where the Kuroshio Current Extension and Gulf Stream strongly influence air-sea exchange and weather activities. In our study, we also use IVT as a proxy to evaluate the skill in representing extreme weather. This information is added to Lines 42-50 in the revised text.

3. Halfway through the paper you write the following: "These errors are not representative of the impact of coupling, nor of SST error in the hindcast, which is shown to be improved in GEPS6 compared to GEPS5 (Lin et al., 2019, Figure 24).". This is startling to say the least. It sounds like you're saying that the comparison between GEPS6 and GEPS5 you are making here is not telling you anything about either coupling nor SST biases because the initialisation of GEPS6 is so different from

that of GEPS5. Doesn't this compromise every single result of this paper? Aren't you trying to exactly assess the impact of coupling or SST errors in the hindcasts? Can you please clarify the exact differences in the initialization between GEPS6 and GEPS5 forecasts and how much these compromise the results? One way to assess what differences are related to the initialization could be to show figures of the day 1 difference, since notable differences in the SSTs/ice at this point should be dominated by the different initialization. All this needs to be discussed in the revised paper.

Reply: We thank the reviewer for the input. We agree that the differences in initialization of SST are a potential challenge. As part of our analysis, we have looked at this issue closely. As shown in Figures 1 and 2, the first day and pentad SST differences are very similar.

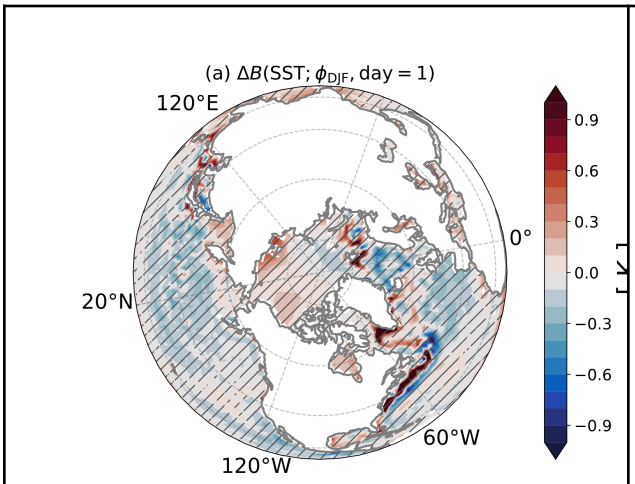


Figure 1: Bias GESP6 minus GEPS5 plot of SST of the first day.

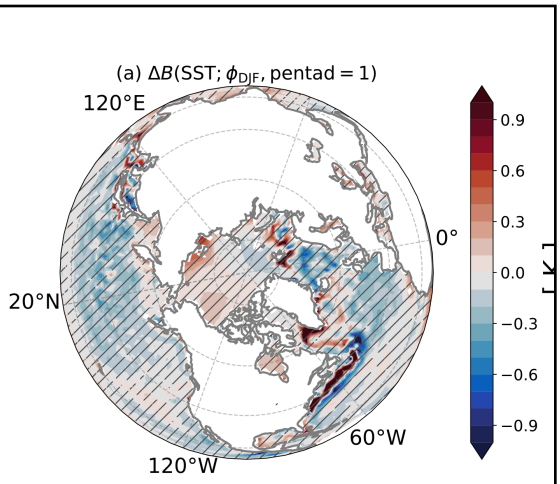


Figure 2 (revised manuscript Figure 1a): Bias GESP6 minus GEPS5 plot of SST of the first pentad.

In the revised text, we have clarified that the first pentad of the atmosphere state represents the initial response to SST initialization differences (Line 152). As stated in the shared reply, in Lines 156-157 we have also added a sentence addressing the bias variance introduced due to such SST difference.

4. You emphasize the importance of the MJO and the sensitivity of MJO forecasts to coupling, but no comment is made about MJO forecasts in GEPS6 versus GEPS5. Has this been looked at previously? How does/might this impact your interpretation of MJO dependent impacts?

Reply: Yes, the MJO prediction skill in GEPS5 and 6 was reported in GEPS6 tech documentation. We have added this information to the introduction (Lines 48-49).

Sensitivity to coupling does not impact our original interpretation. The lag of better MJO prediction in different MJO phases is a reflection of the downstream impact of the MJO. We have added this comment into Lines 181-183.

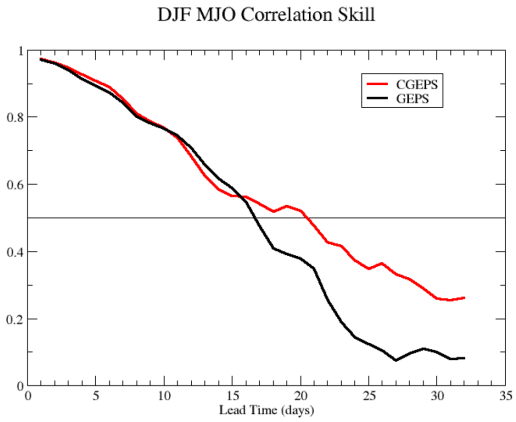


Figure 6. MJO bivariate correlation skill as a function of lead time for the 4-member ensemble mean of the reforecast for GEPS 5.0.0 (black) and GEPS 6.0.0 (red).

Figure 3 (Figure 6 of Lin et al 2019)

MINOR COMMENTS

1. Section 2.3: I don't follow the reasoning here. You say that you can't directly compare the coupled and uncoupled forecasts because the start dates differ, and so you rather compare the two biases instead. However, unless I misunderstood something about the exact computation, this ends up being the same thing: (GEPS6-ERA5)-(GEPS5-ERA5)=GEPS6-GEPS5. So your bias difference plots are just showing GEPS6-GEPS5 anyway. I don't think you can sidestep the problem that the initialization days are different. You just need to mention this as a confounder and discuss how much you think the results depend on it.

Reply:

We thank the reviewer for this input. Yes, as the hindcast was produced operationally every Thursday, the start-date difference is unavoidable. We have added this comment in Lines 91-92: "… In our focus months, December, January, and February, the resulting start times of GEPS6 are exactly one day earlier than those of GEPS5."

As for the concern "(GEPS6-ERA5)-(GEPS5-ERA5)=GEPS6-GEPS5", we need to clarify what was computed is

$$\Delta\beta(t_s,\ t_l) = [GEPS6(t_s - 1\ day,\ t_l) - ERA5(t_s + t_l - 1\ day)] - [GEPS5(t_s,\ t_l) - ERA5(t_s + t_l)]$$

Where $\Delta\beta(t_s,\ t_l)$ is the difference in bias of a start date t_s and lead time l_s. Assuming the validity of the approximation

$$\frac{ERA5(t_s - 1\ day + t_l) - ERA5(t_s + t_l - 1\ day)}{1\ day} \approx \frac{GEPS6(t_s - 1\ day, l) - GEPS6(t_s - 1\ day, l)}{1\ day}$$

because GEPS5 was meant to capture the actual weather. Using this approximation and applying Taylor expansion allows us to express $\Delta\beta$ as

$$\Delta\beta(t,\ l) \approx GEPS6(t_s, t_l) - GEPS5(t_s, t_l),$$

which is what we are looking for. The use of the pentad average window on lead time is an additional layer of insurance to smooth out synoptic-scale fluctuation.

If the reference is not used,

$$\Delta\beta'(t_s, t_l) = GEPS6(t_s - 1\ day,\ t_l) - GEPS5(t_s,\ t_l)$$

This causes problems in quantifying the variance. To see this, rewrite the equation above using Taylor expansion,

$$\Delta\beta'(t_s,\ t_l) \approx GEPS6(t_s,\ t_l) - GEPS5(t_s,\ t_l) - \frac{\partial GEPS6(t_s', t_l)}{\partial t_s'}\Big|_{t_s\ day} \times 1\ day$$

Where the last term is the dependency of bias as a function of start date. This term is large because the patterns of weather anomalies move fast, and therefore, the magnitude of this term is typically the same as the weather anomalies themselves.

In summary, if a reference is not used, the variance due to this term overwhelms our signal of interest. This is why we need reference data ERA5 to remove such a signal.

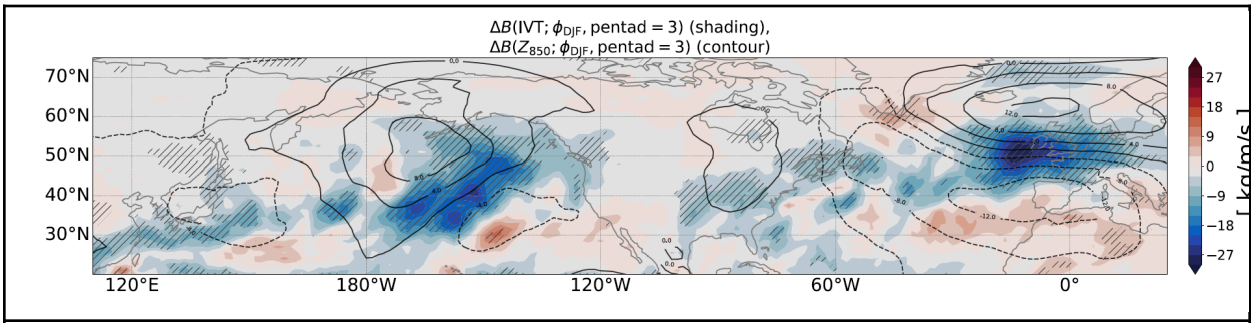
We have put a complete reasoning in the Supplementary Text and direct readers to the text in Lines 128-129.

2. “To test the significance, the degrees of freedom are counted by making the following two assumptions: (a) output from different start times or different ensembles is independent” I guess you mean different ensemble members, not different ensembles. As for the first point, this should be fine as long as the start times are relatively spaced out. Can you comment on the typical distance between start dates? The information is in the supplementary tables but it is convenient if you just state this here for the reader.

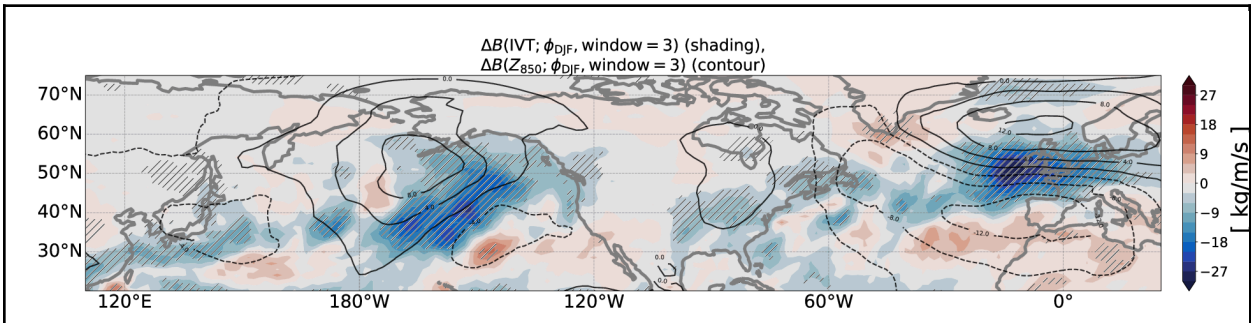
Reply: The typical temporal separation between start dates is about a week. And yes, we should have written “different ensemble members”. We have corrected this in Line 121. We have also clarified the temporal separation between start times in Line 92: “…and start dates are spaced by 7 days”.

3. Figures 2/3: Can you make the continents more visible? Coastlines blend in with contour lines, making it hard to distinguish the two.

Reply: We have improved the figures as suggested. Here is the comparison of old v.s. new Figure 2:



Old Figure 2



New Figure 2

4. L140: You should include a few lines on how the shift in the Gulf Stream is very likely related to the inability of NEMO at ¼ degree to resolve the Gulf Stream properly, and cite some references for the role of model resolution. I don’t know as much about the Kuroshio current, but I’m sure model biases in this current, and likely origins of such biases, have been looked at in past studies, so it would be good to discuss these briefly as well. Alternatively, you could add this discussion to your section 4, but if so, please mention here that you will discuss these biases further in section 4.

Reply:

Thank you for this input. We have added extra sentences and citations into the discussion in Lines 232-236: "… While the resolution of the 0.25° ocean model used in GEPS6 is sufficient to resolve mesoscale eddies (0.5°– 2° or 50–200 km) (An et al., 2023), Chassignet and Xu (2017) suggest that much finer resolution (less than $1/12^{\circ}$ or 8 km) is required to resolve the smaller eddies to obtain the observed magnitude of eddy kinetic energy in boundary currents, and therefore adequately resolve the positions of boundary separation and eastward turns for both the Kuroshio and Gulf Stream currents."

5. L150: The link between the Aleutian and Icelandic lows is known and documented, see e.g. Honda et al. (2001): https://doi.org/10.1175/1520-0442(2001)014<1029:ISBTAA>2.0.CO;2 Please add some references here.

Reply: We have added this reference to the paper in Lines 164-166: "This is consistent with reanalysis findings by Honda et al. (2001), who linked the influence of the Aleutian Low on the Icelandic Low at subseasonal time scales. These linkages have also been widely reported at seasonal and longer timescales (Li et al., 2024, and reference within)."

6. L171/172: "indicating that the coupling results in a colder SST" Can you add a comment on why this might be? This comment might be related to the above comment about past literature on Kuroshio current biases in models.

Reply: We thank the reviewer for this input. In the revised text, we have added a comment in Lines 217-218 to clarify this point: "… −0.2 K in the first pentad due to difference in SST initialization strategy, and potential issues in model resolution (see Section 4)."

7. L208: Figure 3d-f should presumably refer to Figure 1d-f.

Reply: Thank you for catching this. We have fixed it (now in Line 239).

8. L211: "Future numerical studies are needed to gain a deeper understanding." Figure 1f and 1i show an NAO pattern in the Euro-Atlantic. The relationship between changes in the Gulf Stream and changes in the NAO have been investigated in many past studies, see e.g. this paper and references therein: https://doi.org/10.1029/2025GL117228

More pragmatically, the NAO is the dominant mode of variability there so if you change the SSTs in this region then the atmospheric change is very likely going to project onto the NAO. Please add some comments on this, especially on the past literature.

Reply: We have added comments and references in Lines 241-243: “This aligns with reanalysis studies showing that Gulf Stream variability leads the North Atlantic Oscillation (NAO) by a month through its strong temperature gradient that impacts boundary processes during cyclogenesis (Parfitt and Kwon, 2020; Chakravorty et al., 2024; Alsepan and Parfitt, 2025). “

9. L216: “We also notice a possible teleconnection from the Aleutian Low through the Arctic into Icelandic Low via a Rossby wave train.” Since this teleconnection is known (see above), you should rephrase to rather say that the changes to the Aleutian Low affect the Icelandic Low via a Rossby wave train, and then cite Honda again.

Reply: We have amended the sentence and have added the citation. Now Lines 250-252 read: “We also note how the biases in the Aleutian Low due to differences in SST initial conditions can subsequently impact the Icelandic Low via teleconnections, as is consistent with reanalysis studies (Honda et al., 2001; Li et al., 2024).”

10. L233/234: “Second, there is a need for more physical understanding of how two-way coupling produces better air–sea fluxes.” There are some classic relevant studies on this. Most notably, Barusgli and Battisti (1998) needs to be mentioned here: https://doi.org/10.1175/1520-0469(1998)055<0477:TBEOAO>2.0.CO;2

In this paper they clearly explain the effect of coupling versus no coupling on heat flux and surface temperature variability. In particular, the low frequency variability in surface temperature (and, I believe, heat fluxes) will be wrong in uncoupled models due to the excess thermal damping effect they explain there. This is fundamentally related to the fact that the ocean acts as an infinite sink/source of energy in an uncoupled simulation. It seems plausible that changes in the latent heat flux bias variance you see could be related to this. You don’t necessarily need to demonstrate this decisively, but some comments at least are necessary.

Reply: The study mentioned seems to reach conclusions for timescales longer than 100 days, as can be seen from their Figure 4:

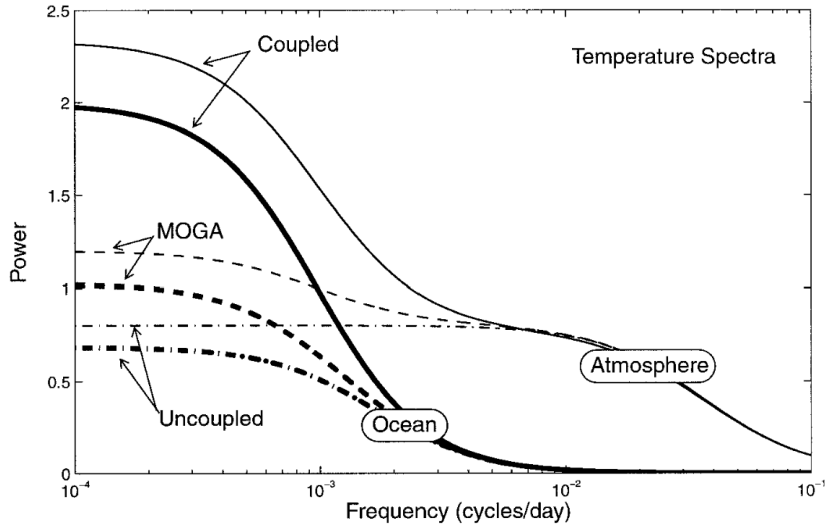


FIG. 4. Power spectra of atmosphere and ocean temperature for the coupled, MOGA, and uncoupled cases. The standard parameters (see Table 1) are used.

Their results show that air-sea coupling is important for a statistical model at timescales longer than 100 days ($10^{-2}\frac{cycles}{day}$). In contrast, we focus on timescales shorter than 15 days (right-most part of the figure). Therefore, in our assessment, the Barusgli and Battisti (1998) results are not directly applicable to this study.

Citation: https://doi.org/10.5194/egusphere-2025-4142-RC1

Reply to Reviewer 2's Comments

Reviewer #2

The paper describes the differences between seasonal hindcasts produced using the GEPS5 (uncoupled) and GEPS6 (coupled) systems. The authors identify some interesting coupled feedbacks and the paper is generally well written and presented. However, there are some significant issues that need to be addressed in order to make it acceptable for publication.

Major comments

1) The biggest issue is that the initialisation of the SST between the two systems is so different and uses different start dates. This makes it very difficult to be certain that the changes highlighted are due to coupling and not due to the initialisation method. You need to distinguish between the impact of SST differences due to initialisation and SST differences due to running a coupled model.

Reply: We thank the reviewer for this input. In the revised manuscript, we have followed Reviewer 1's suggestion to frame the paper as the impact of replacing persistent SST anomalies with a dynamical ocean model. Throughout the manuscript, we have clarified that the initialization of SST causes the initial perturbation, and that interactive air-sea coupling explains certain aspects of the evolution of the perturbation. The response of the atmosphere to different SST initialization strategies is reflected in the first pentad. It will be our future work to design experiments to remove the impact of initial conditions.

2) The authors show the differences between the biases of the two systems but they don't state whether the changes are an improvement or not. i.e. where GEPS6 is colder than GEPS5, is GEPS5 warm compared to observations and therefore GEPS6 is better or is GEPS6 too cold. I note in section 3.4 you suggest that GEPS6 is better and cite Lin et al 2019. This appears to be an internal document at ECCC that I have been unable to find online. Ideally, the bibliography should state how to get this document or you should include the SST bias plots in the current paper. It would be better to discuss this as part of section 3.1 rather than where you cite the Lin report in section 3.4.

Reply: GEPS6 improves the tropical SST forecast as documented by Lin et al (2019). However, SST bias variance in GEPS6 is worse than GEPS5 if measured globally or in either hemisphere (Figure 1 below). Overall, GEPS6 adds about $0.05\ K^2$ error variance to the global ocean. We have added this information in Lines 156-157: "In the Northern

Hemisphere, the difference in SST initialization strategy introduces a bias variance of about 0.05 K^2 (Supplementary Figure S2)."

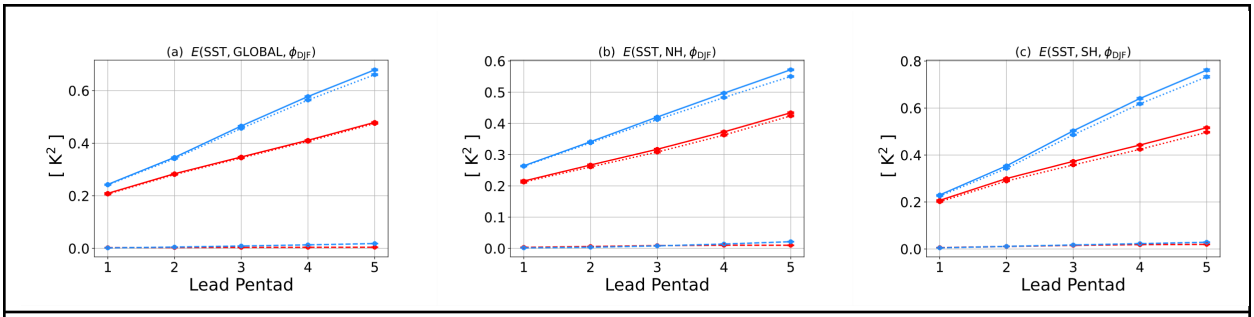


Figure 1 (Supplementary Figure 2 in the revised manuscript): Bias variance (E) analysis of global sea surface temperature (SST) computed from Global Ensemble Prediction System (GEPS) version 5 (GEPS5, red) and GEPS version 6 (GEPS6, blue) during December-January-February of lead pentads 1 to 5 in hindcast years 1998--2017. The regions presented are (a) global, (b) Northern Hemisphere, and (c) Southern Hemisphere oceans. The decompositions of E into mean ($\overline{E}$, dashed) and patterned ($\widetilde{E}$, dotted) variances are added. Error bars represent the standard error.

We have also provided the hyperlink for Lin et al (2019) in the Supplementary material and have noted this for readers in Lines 46-47.

Minor Comments:

1. Section 3.3: You identify the positive feedback between the cold SST and Z850. Given my comment above about initialisation, can you be sure that the colder SST in the coupled model is due to being coupled or due to the initialisation strategy. Also, how far into the forecasts does this positive feedback persist? Presumably, it must reach an equilibrium at some point.

We thank the reviewer for the input. Indeed, we cannot rule out the impact of initialization. A more appropriate interpretation is that initialization imposes an initial perturbation on the system, and air-sea coupling damps or amplifies that signal. Judging from our plots extending into pentad 5, the weakened Aleutian Low anomaly dissipates after pentad 3. But it is tricky to reach a conclusion on the persistence of the feedback due to the difference in the SST initialization. A cleaner experiment is outside the scope of the current study but will be conducted in the future, likely using a regional model to remove remote influences. We have added a comment in Lines 225-227 in the revised text: "Given differences in the initial SST and the use of a global model, further isolation of this feedback is beyond the ability of the present framework. Nonetheless, the results underscore the potential of future regional modeling studies to more directly quantify the strength of the feedback."

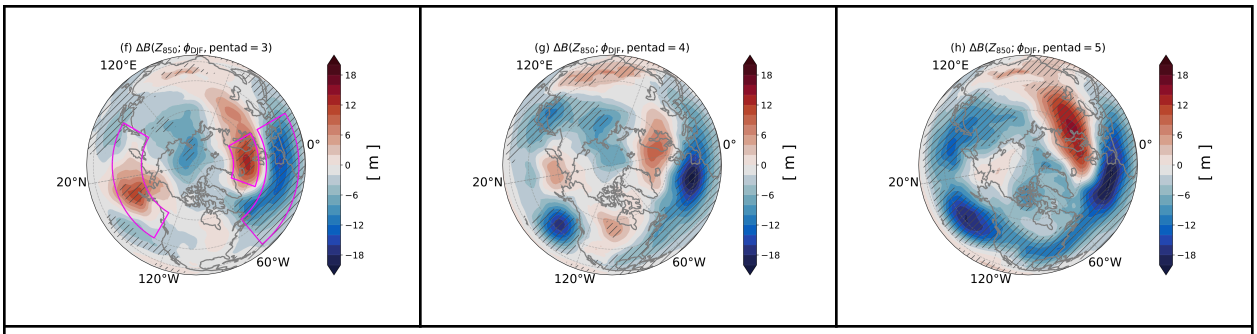
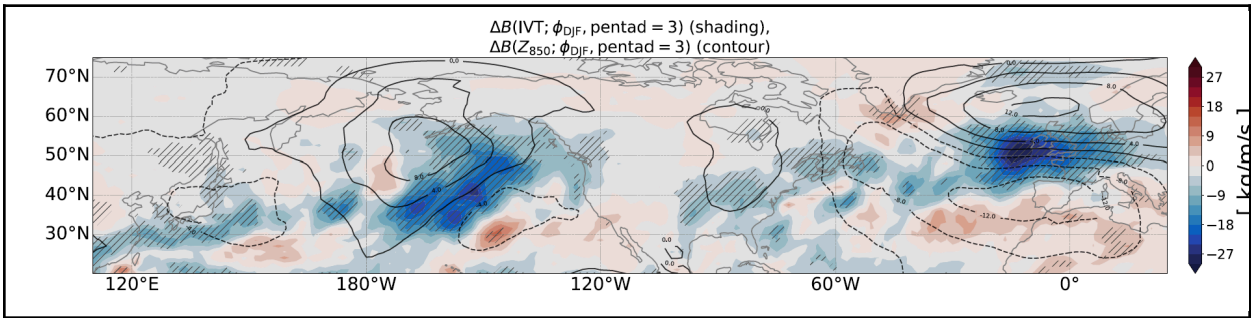


Figure 2: Bias change of Z850 from pentad 3 to 5.

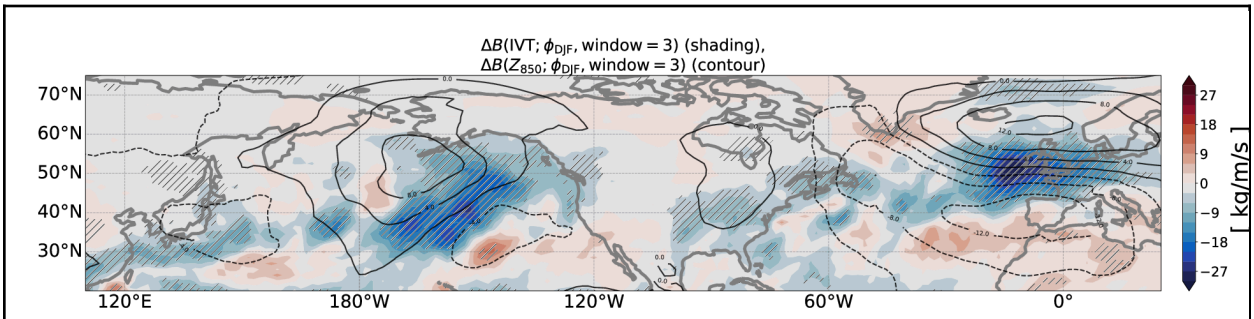
2. Figure 3: It's very hard to see the location of the continents on this figure.

Reply: Thank you for the input. We have thickened the coastlines in the updated manuscript. This automatically applies to Figure 2 as well. Here is a comparison:

Here is the comparison of old v.s. new Figure 2:



Old Figure 2



New Figure 2

3. Figure 6: You include this in the paper but there is no reference to it at any point in the text. I think some discussion of this figure would be interesting to understand whether the impact of the dynamical coupled model is different according to the MJO phase. Perhaps this is what you are referring to in your conclusions at line 226 ("The

IVT improvement is also more significant when the MJO is active"). If so, this should be introduced and discussed before this point in the paper.

Reply: We thank the reviewer for pointing this out. We have added a discussion of this figure and topic to Lines 179-183: "Over the North Pacific, the IVT forecast is improved when the MJO is active. Figure 4a–c shows the composited bias variance of IVT over the Kuroshio Current region. The first three pentads of non-MJO cases do not show significant differences, while the MJO phases 1–4 and 5–8 show better IVT forecasts starting in pentads 3 and 2, respectively. The lag in the improvement in MJO phases 1–4 by one pentad is reasonable because MJO convection is located over the Indian Ocean during phases 1–4, and the signal takes some time to propagate into the Pacific."

Please note that the old Figure 6 is now Figure 4 in the revised manuscript.

4. Lines 229-231: This sentence confused me. You haven't discussed the parameterizations used for calculating the air-sea fluxes, just compared the results with different SST patterns. Are you suggesting that your results motivate further work to understand why the latent heat flux bias variance is reduced even when the SST bias variance is increased?

Reply: We thank the reviewer for raising this point. We have decided to replace the sentence identified by the reviewer with improved sentences in Lines 263-268, which point out the direction of improvement for each basin: "This basin-dependent behavior implies different limiting factors in the North Pacific and Gulf Stream regions. In the North Pacific, the quality of the initial SST is high enough that improvements can be made through better air–sea interaction, such as higher-order turbulent mixing schemes or the inclusion of a wave model (Sauvage et al. 2023). In the North Atlantic, the initial Gulf Stream SST bias remains large such that improving the air–sea interaction will not yield significantly improved forecasts, unless better air–sea interaction leads to a higher quality initial assimilated SST."