We thank the reviewer for the positive assessment of our manuscript and for the constructive comments and suggestions. We have carefully considered all points raised and revised the manuscript accordingly. Our detailed responses are provided below.

1) l. 28, when discussing the potential for S2S prediction using MLWP models, some references are missing to reflect what has been done already. For instance, the two following papers are relevant references to include as they discuss and demonstrate the advance of S2S forecast skill using these models.

Weyn, J. A., Durran, D. R., Caruana, R., & Cresswell-Clay, N. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. Journal of Advances in Modeling Earth Systems, 13(7), e2021MS002502. https://doi.org/10.1029/2021ms002502

Chen, L., Zhong, X., Li, H., Wu, J., Lu, B., Chen, D., et al. (2024). A machine learning model that outperforms conventional global subseasonal forecast models. Nature Communications, 15(1), 6425. https://doi.org/10.1038/s41467-024-50714-1

Thank you for suggesting the important reference to S2S forecasts, now 2 papers have been added in the paper. (See l. 23)

The growing number of studies demonstrating skillful medium-range MLWP performance suggests that these systems also offer a promising path for advancing S2S forecasting (Weyn et al., 2021; Chen et al., 2024), even as several important challenges remain. For instance, data-driven models like Pangu-Weather (Bi et al., 2022) and hybrid approaches such as NeuralGCM (Kochkov et al., 2024) have demonstrated skill comparable to state-of-the-art NWP models in the medium range.

2) l. 85 : "Sea surface temperature are prescribed from ERA5." Can you detail here? Do you maintain SST anomalies from initialization (persistent SST anomalies)?

For NeuralGCM, with a dynamical core, we can prescribe SST from its forcings. We have the sensitivity test on a persistence forcings (keep SST and sea ice concentration constant during forecasting), and also dynamic forcings(pass the daily updated SST and sea ice concentration excluding their diurnal cycle to forcings) from initialization, the results in the end doesn't significantly change the results. In the website of NeuralGCM, there is a clear description in "**Advancing in time**" the section of "Deep-dive into trained models" (https://neuralgcm.readthedocs.io/en/latest/deepdive_into_models.html#)

3) l. 91: "This leads to its significantly lower computational resource requirements compared to the other two models of this study." Could you give an rough estimate of each MLWP model's computational cost here, relative to UFS?

For UFS, running relaxation experiment on subseasonal time scale takes a few days while MLWP models only takes a few hours.

4) Section 2.3 : it sounds like the daily anomalies for the models are calculated from the ERA5 daily climatology. Ideally the model anomalies should be calculated using the model daily climatology, but this requires a set of hindcasts over a sufficient long period. I do not think that

using model climatology would significantly change the results, but this should be mentioned for transparency.

Thanks for pointing it out. What you stated here is correct, it is quite difficult for MLWP models to get model climatology. So we set up the model replay experiments. Model replay is we relaxed the whole globe (which you would consider model bias during the relaxation and forecasting) and this is served as verification data, and this is almost the same as ERA5. And this model replay can also examine if relaxation functionally works well in the MLWP model and didn't introduce large model bias to S2S forecasts.

5) l. 125, it is unclear what the "model replay" experiment is used for in the study.

Thanks for your feedback. The sentence has been modified correspondingly (See l. 126).

The four types of relaxation experiments are Control (CRL), narrow tropical relaxation (NTR, relaxing from 20°S to 20°N including the tapering region), wide tropical relaxation (WTR, relaxing from 30°S to 30°N including the tapering region). Model replay is that we applied relaxation globally and vertically to include model bias and served as verification dataset as ERA5. More detailed information are available in Table 1. WTR is designed to assess the overall impact of the entire tropics, whereas NTR focuses more strictly on the deep tropics. In UFS, horizontal wind components, geopotential, specific humidity and temperature are nudged (Table 2). In Pangu-Weather, variables at all 13 pressure levels are nudged during model integration. All surface level variables, such as 2-m temperature, are excluded from relaxation in Pangu-Weather. In NeuralGCM, all variables except geopotential are relaxed along vertical layers between boundary layer and tropopause, including specific cloud ice and liquid water content. The relaxation is applied every 24 hours in both MLWP models.

6) Section 3.1: the December case study has also been highlighted in our recent paper (Peings et al. 2026), as a window of opportunity for S2S forecasting. The three models used in our study (two MLWP models and the ECMWF S2S model) exhibit good prediction skill for this period at week 2 as shown in the paper, but we also found good skill for week 3 and more generally for the week 2-4 window. We also preformed a sensitivity study with one of the MLWP model to demonstrate that the skill was coming from the tropics. I think this paper is worth being cited because it aligns with the result presented here.

Peings, Y., Dong, C., Mahesh, A., Pritchard, M., Collins, W., & Magnusdottir, G. (2026). Subseasonal forecasting and MJO teleconnections in machine learning weather prediction models. Journal of Geophysical Research: Atmospheres, 131, e2025JD044910. https://doi.org/10.1029/2025JD044910

Thanks for providing your helpful and valuable findings in your paper, it has been cited in the manuscript. (See l. 192)

event. The similarity between the WTR and NTR forecasts for all of the models suggests that a better representation of the tropics would have improved the subseasonal forecast skill for this event. This finding is also align with the latest study of Peings et al. (2026), where the improved S2S forecast skill comes from the tropics in a MLWP model through sensitivity tests.

To further understand how the relaxation in the tropics impacts the extratropical Rossby wave forcing, we analyze the RWS (Section 2.4.3) at 200 hPa averaged from 23–30 December 2022 (during week 2; Fig. 3), which is one week earlier than

7) The section about the physical mechanism leading to more skillful predictions for the two case studies would benefit from being developed. The RWS anomalies of Fig. 3 and Fig. 5 are

noisy and they are not very explicit. I think it would be interesting to see how they bridge the tropics with the extratropics. I.e., showing the Rossby wave associated with it, maybe at different lead times (week 1, 2 and 3) to show its development. You could also show how the deep convection anomalies in the tropics differ in CRL versus NTR in function of time, maybe using a Hovmoller plot (time in function of longitude) which would reveal how MJO propagation changes with nudging and makes for a more accurate teleconnection. The paper only includes 6 figures so there is room for a couple figures further detailing the tropics-extratropics teleconnection leading to improved skill in the North Pacific/North America sector (especially for the December case).

Thank you for your nice suggestions. We have plotted PV200 as proxy for 3 models in CRL vs NTR in the tropcis (-10,10) in the Hovmoller plot to see tropical convection and MJO propagation for both 2 cases and will update those plots in the manuscript.

8) In conclusion, when stating that "However, drawing more definitive conclusions will require a systematic evaluation over multiple years and similar events to assess the generalization of these results", it should be mentioned that a systematic evaluation of the S2S forecast skill for the North Pacific/Western North America region has been done for NeuralGCM (Peings et al. 2026). The study shows that two MLWP models (SFNO-HENS and NeuralGCM) exhibit comparable S2S skill to ECWMF for the case of the MJO and North Pacific atmospheric patterns during the October-March season.

Thank you for pointing out this important study. We have revised the conclusion to acknowledge the systematic evaluation in your study and to better contextualize our statement regarding the need for further assessments (See l. 292).

290 – The z500 forecast skill of the CRL experiment with Pangu-Weather and NeuralGCM exceeds that of the UFS. These findings underscore the promise of data-driven models in subseasonal forecasting, particularly given their lower computational costs. Recent work by Peings et al. (2026) has also provided a systematic evaluation of S2S forecast skill over the North Pacific/Western North America region, showing that two MLWP models (SFNO-HENS and NeuralGCM) exhibit skill comparable to ECMWF for MJO-related and North Pacific atmospheric patterns during the October–March season.
295 Nevertheless, further systematic evaluations across multiple years and a broader range of events are still needed to fully assess the robustness and generalization of these results.

9) l. 296: "This suggests that a better representation of the tropical atmospheric state in the models would have improved the prediction of this particular event". The conclusion would benefit from a discussion of how the MLWP models have the potential to improve prediction skill in the tropics, and consequently in the mid-latitudes (if they do).

Do the authors anticipate that S2S forecast skill will improve with future developments in both traditional dynamical models and machine-learning weather prediction (MLWP) systems? Or does the current similarity in S2S skill between MLWP models and GCMs indicate an intrinsic predictability limit of the climate system that may be difficult to surpass?

Nudging simulations such as those presented in the paper are valuable for investigating mechanisms and tracing potential sources of predictability for specific events. However, do we realistically expect S2S forecasts in the tropics to become sufficiently accurate to substantially improve prediction skill in the mid-latitudes? I know that is the million-dollar question bit it would be worthwhile to address it in the conclusion to place the results in a broader predictability context.

We thank the reviewer for this valuable comment. There are still open and interesting questions remained. In the paper of Judt, 2020 (https://doi.org/10.1175/JAS-D-19-0116.1) also suggested the tropics have longer predictability than the middle latitudes and polar regions. Whereas tropics has more than 20 days, and middle latitudes and polar regions have a little over 2 weeks. We do see there is predictability limit but still a lot of space to improve. And we have added a discussion in the conclusion (lines 318–323) addressing there are MLWP models could improve predictability (such as AIFS), including recent developments in loss functions that could potentially enhance tropical variance representation.