Referee 1 (Peings):

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 and highlighted in orange font color.

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 references to ML-based S2S forecasts. The two papers have been added in the paper (See l. 23).

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, SSTs can be prescribed as a continuous external forcing which is taken here from ERA5 at 24 hourly resolution. Of course this leads to an apparent advantage of NeuralGCM over UFS for which forecasts are done with a coupled model. We tested the potential sensitivity of the NeuralGCM results with forecasts using persistent SST forcings (SST and sea ice concentration at forecast initialization time are kept constant during forecasting), and also dynamic forcings (daily updated SST and sea ice concentrations). The results show only minor sensitivities. In the revised manuscript, we now provide additional information as also requested by the second reviewer (ll. 86-91).

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, for example, NeuralGCM run 10 ensemble forecast for 30 days with 38 minutes , 1548412 Joule / 430.114 Watt/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 absolutely correct, it is quite difficult for MLWP models to get model climatology which is not falling into the training period. To estimate the potential effect of model behaviour on the results, we conducted the model replay experiments. During these experiments, atmospheric fields are nudged globally and serve as verification data. It turned out that the replay experiments are almost identical to ERA5 so that we decide to not run a large set of hindcasts. We mention this potential limitation of our approach in the revised manuscript. (ll. 107-110)

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. 137-141).

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 ll. 221-225)

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 nice suggestions to help deepen the understanding of the tropics-extratropics teleconnection. We now include Hovmoeller plots of velocity potential anomaly at 200 hPa in the tropics as a proxy of tropical convective activity (Fig.4&7, ll. 244-264, ll. 314-326). Our analysis reveals that convective activity is overestimated in NeuralGCM and Pangu-Weather

which is consistent with UFS. The tropical relaxation reduces the velocity potential/the convective activity. The MJO propagation does not seem to be affected substantially.

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 ll. 352-356).

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. Indeed this is the million-dollar questions and interesting questions remain. The seminal paper by Judt, 2020 (https://doi.org/10.1175/JAS-D-19-0116.1) suggests that the tropics exhibit a substantially longer intrinsic predictability than the middle latitudes and polar regions. Whereas tropics have an intrinsic predictability limit of more than 20 days, middle latitudes and polar regions have a little over 2 weeks. So, there is a intrinsic predictability limit but following their argument there is still a lot of space to improve. That MLWP models do better exploit the intrinsic predictability than NWP models is potentially also shown by very recent work of Polichtchouk et al. 2026. They show that the forecast skill horizon of an NWP model in the tropics can be extended by as much as two days if it is nudged toward an MLWP model. We have added a corresponding discussion in the conclusion (lines 381–394).

Referee 2:

The study discusses how constraining the atmospheric state in the tropics improves sub-seasonal forecasts of two extreme precipitation events in western North America in the winter of 2022/2023 in two machine-learning weather prediction (MLWP) models. It complements an earlier study by some of the authors, where an equivalent relaxation experiment was performed in a physics-based weather prediction model. It is found that the response to the tropical constraint in the MLWP models is similar to that in the physics-based model, although somewhat weaker owing to a stronger baseline performance. An analysis of Rossby wave sources indicates that the MLWP models simulate the tropical-extratropical teleconnections contributing to the extreme events in a physically consistent way. The authors emphasise the general point that such relaxation experiments are a useful diagnostic tool to understand and improve sub-seasonal predictions.

The paper is a useful contribution to the field of MLWP model evaluation and merits publication. Since it discusses two very specific case studies, it lacks the generality that the title suggest, but on the other hand there is value in a detailed assessment of how MLWP forecasts represent tropical-extratropical teleconnections for specific mid-latitude extreme precipitation events. The presentation is mostly clear and concise, but I would recommend some clarifications and revisions, as well as adding some further analysis and discussion - see the comments below.
We thank the reviewer for their careful reading of the manuscript and for their constructive comments. These helped to improve the clarity and scientific quality of our manuscript. In the following, we respond to each comment. Our responses are highlighted in orange font color.

1. - Would it be worth investigating the reasons for a different importance of tropical forcing between the two cases a bit more, e.g. by looking into origin and propagation of forecast errors over time? The RWS diagnostic only works for tropical sources, but it would be good to quantify mid- and high-latitude contributions.
We are grateful for this suggestion but are somewhat uncertain whether the question refers to the different importance of tropical forcing in terms of dynamics or predictability. In the following, we focus on differences in terms of predictability as this is indeed striking. In order to illustrate the models' behaviour over time, we now include Hovmoeller diagrams of upper-tropospheric velocity potential as a function of lead time. ERA5 reveals that both events were associated with the MJO-characteristic dipole of suppressed and enhanced convection centered around 90°E. For the February case, the MJO-related convection was slightly stronger than for the December-case. For the latter, NWP as well as MLWP models overestimated the convective activity already at lead time of 10 days. In contrast, models exhibited a better representation of the MJO and the associated velocity potential for the February case already without any tropical nudging. Hence, we conclude that a major reason for the different importance of tropical convection for the predictability is due to the representation of the MJO-related convection. To quantify, mid- and high-latitude contributions would require carefully designed relaxation experiments as it would be meaningless to evaluate forecasts in regions which have been nudged. These, experiments however, would substantially expand the study and are thus out of scope. In the revised paper, we argue as outlined above and hope to have answered your question suitably.

2. - The title is too general for what is being presented. I suggest to start from the title of the reference study by Moore et al. ("Impacts of tropical forecast errors on weeks 3–4 extreme precipitation predictions over California during winter 2022–23") and modify this to reflect the new aspect of relaxing MLWP models

Thank you very much for suggesting a revised title. In the revised paper, the title reads: “Impacts of Tropical Forecast Errors on two Extreme Precipitation Events: Insights from Relaxation Experiments using Machine-Learning Weather Prediction Models”

3. - l. 7: the fact that only tropical relaxation is considered should be mentioned earlier than this, potentially first sentence of the abstract
Thank you for this suggestions. This information has been added in the abstract in l. 3.

4. - l. 73: "6-hour forecast increment" - I suspect you are referring to the output available, not to the model time step (which would be hard to believe). Please clarify.
Thanks for pointing this out, it is indeed referring to the output available. We have clarified it in the manuscript in l. 74.

5. - l. 85: The fact that NeuralGCM uses "perfect" SST prescribed from ERA5 strikes me as important. It means that the NeuralGCM setup could not issue real-time forecasts, and it should have an unfair advantage over the other models considered. What is your view, maybe you can add some discussion or analysis on this?
Thank you for raising this valid point. NeuralGCM has an advantage over UFS as this is run with a coupled ocean. As the current version of NeuralGCM does not allow ocean coupling, we assessed the potential effect by runing NeuralGCM with fixed SST taken at initialization time from ERA5. We found that the difference between runs with fixed and prescribed SST is not able to explain the difference found in skill between NeuralGCM and UFS for these two events. So overall, the different strategies with regard to ocean coupling complicate the fairness of comparison among the three models, but it does not invalidate results. We address this aspect in the revised paper in ll. 86-91.

6. - ll. 89-90 (..., Pangu-Weather is ..."): Please specify which Pangu model you are actually using for your inference relaxation study - is it the on with a 24h time step?
Yes, here we only use the Pangu model with the 24h time step to run relaxation experiment. We have added this information in the manuscript. See ll. 97-98.

7. - ll. 90ff. ("Overall the model is..."): I don't understand this, please rephrase. With "auto-regressive during training" I assume you are referring to rolling out for more than one model time step during training and minimizing the loss computed from the rolled-out errors. This is indeed more costly during the training, but has no impact on inference cost. The real reason that Pangu is cheapest among the models you are considering is probably that its does not need to run a GCM dynamical core (expensive, both UFS and NeuralGCM have it).
Thank you for pointing out the ambiguity in our wording. We agree that autoregressive training (i.e., multi-step rollout during training) primarily affects training cost and does not determine inference cost. We have revised the manuscript to clarify this point and removed the potentially misleading statement. See from ll. 96-97

8. - l. 96: Why did you choose to compute the climatology over this long period? Can you please check whether substantial trends are present for the variables you are considering? If this is the case, there is the risk that anomaly correlations presented are inappropriately dominated by these trends.
Thank you very much for this comment as this made us realize that we had provided incorrect information. The climatology is calculated for the period 1990-2019 and not 1970-2019. Still, we

agree that if there were substantial trends for the variables this would affect the absolute values of the ACC. However, as we are using the same period for all experiments and are mostly concerned about the differences between models and between CRL, NTR, and WTR, our main conclusion will not be affected by the trend.

9. - Table 1: The "Replay to ERA5" experiment is never used in the manuscript. Why? Please either remove the reference to this experiment, or use it when discussing the results.
Indeed, the replay to ERA5 is not shown in the manuscript. The purpose of the replay was to investigate potential affects of model biases and to examine that the nudging in data-driven models yielded sensible results at all. When comparing the replay to ERA5 and ERA5 itself, we found that these were almost identical with regard to the variables analysed here so that we included only ERA5. For the sake of completeness, we now show the replay figures in the supplemental material. (Figure. A2)

10. - ll. 115ff.: Did you test the sensitivity to the width of the tapering region, or to the functional shape of the transition? If yes, a comment on that would be helpful.
Thank you for detailed suggestion. We did not perform sensitivity tests in this direction as we aimed for a setup identical to the one of Moore et al. (2026) and as the hyperbolic tangent function has been used successfully in previous studies (e.g., Magnusson 2017).

11. - ll. 119f. ("Relaxed region is corrected by 100% at each time step"): You say on line 108 that the relaxation "gently steers the model state", which is inconsistent with a 100% replacement of the model forecast by the reference. Maybe worth clarifying this on line 108 and elsewhere.
Thanks for pointing out this inconsistency. We have removed the word “gently” from the description.

12. - ll. 127 - 131: Can you please explain the motivation or justification for relaxing different variables for different models? One might argue that this makes the experiments less comparable.
Thank you very much for this constructive comment. Our intention is to keep the experiments as comparable as possible across the three models. Therefore, we use the same initialization and relaxation variables that are common among the models, following the variables nudged in the UFS configuration whenever possible.
However, the prognostic variable sets differ among the models. For example, Pangu-Weather includes surface variables such as mean sea level pressure that are not available in NeuralGCM, while NeuralGCM includes cloud liquid and ice water content that are not present in Pangu-Weather. As a result, the relaxation variables cannot be made fully identical across the models. We performed sensitivity tests for NeuralGCM and found that relaxing geopotential introduces large negative forecast errors. Therefore, relaxation of geopotential was excluded for NeuralGCM in the final configuration. We have clarified this motivation and experimental design in the revised manuscript. (ll.143-146)

13. - ll. 148-151: Please elaborate how you compute and interpret anomaly correlation, it is left a bit vague. As I understand it, this is the pattern correlation between the verification anomaly in Fig 1 and forecast anomalies in Figs 2 & 4. – correct?
Thank you for this helpful comment. Your understanding is absolutely right. We have clarified this description in the revised manuscript (ll. 175–179).

14. - l. 172: I would not call this a forecast bust. Larger errors are expected for any extreme event occurring in the observations when forecasts have modest levels of skill and tend to predict climatology.
Thanks for giving the suggestion on wording. We have rephrased the sentence with a lighter statement and have placed it in the first paragraph of Section 3.1 (See ll.184-185).

15. - Figure 1: The green lines are really hard to see - is it worth making separate panels?
Thanks for your feedback, now we have updated the corresponding figures in the manuscript and show the 40 g kg^{-1} m s^{-1} isoline for water vapour transport in black and bold.

16. - l 181: please define how the water vapour flux is computed (I assume you are showing the magnitude of the vector quantity). Also please add some discussion on why you do not use precipitation directly and what whether this constitute a caveat of the study. It might be worth citing Lavers et al., Weather and Forecasting (2017), https://doi.org/10.1175/WAF-D-17-0073.1 in this context.
We thank the reviewer for this valuable suggestion. The MLWP models used here do not directly predict precipitation, so that we can only use a proxy or related parameter. Since the three models have additionally different vertical levels, we decided to not show integrated water vapour transport but 850-hPa moisture transport. In the manuscript we now explicitly define how the water vapour flux is computed and have cited the paper suggested (See ll.153-159).

17. - Figure 2 caption: Looks to me like the bold green line is at 40 not at 20.
This was indeed inconsistent between panels. We have revised the figures also following your earlier comment.

18. - Figure 3: this is extremely hard to see, even on a very large screen. Please revise to have fewer panels or less details in each.
Thank you for constructive feedback on Figure 3. Now we recreated the figure with less information and hope that essential features are better visible.

19. - l. 235 ("This similarity suggests..."): OK but what does this mean? That only sources in the deep tropics matter?
We thank the reviewer for this clarification. We have revised the text (ll. 292–294) to clarify the interpretation of the similarity between the WTR and NTR experiments.

20. - l. 257: Could this also be because NeuralGCM sees the observed SST (see one of my earlier comments)?
Thank reviewer for raising the good question. The sensitivity test with a fixed SST in Neuralgcm has shown a negligible difference, so the observed SST seems to have little impact on this event.

21. - l. 269: This is a trivial result - any relaxation of ensembles towards a common reference state will reduce the ensemble spread
We thank reviewer's clarification. We have rephrased the statement in the ll. 330-331. It now reads "The substantially reduced range of ACC values between the different ensemble members can be attributed to the reduced variability in the tropics through tropical relaxation.".

22. - Figure 6: Seeing Z500 anomaly correlations of close to 1 for almost every single ensemble

members at 3-4 weeks lead time makes me wonder whether the ACC you compute is a discerning enough metric. Can you please show the same plot for MAE? A reader could conclude from the extremely high ACC for most ensemble members that there is near-perfect deterministic forecast skill in the sub-seasonal range for these events, which I would be sceptical about even with strong impact from tropical sources of predictability.
Thank you for your question, now we have provided the figure in Appendix Figure A3.
23. - l. 276: Can you discuss a bit more why there is less impact for the February event? Given the flow pattern, a stronger impact of mid- or high-latitude dynamics is plausible (see also the Moore et al. study)
Thank you for the helpful suggestions. The weaker impact of the relaxation in the February event may be related to differences in the large-scale tropical variability at initialization. In Case 1, the MJO is not active at forecast initialization time, whereas the MJO is already active in phase 3 in Case 2. When the MJO is active, it represents a coherent and predictable large-scale tropical signal that can already be well represented in subseasonal forecasts. As a result, the February initialization may contain higher predictability in the tropics, leading to the better forecasts of the large-scale patterns in the CRL in Case 2, compared with Case 1.
Because the tropical variability is already better constrained by the initial conditions in this case, the additional relaxation has a smaller effect on the forecast evolution. In contrast, when the MJO signal is weak or absent at initialization, as in Case 1, tropical variability may be less predictable and more sensitive to model errors. Under these conditions, the relaxation can have a larger influence by helping to better capture the variability in tropics.
This clarification has been updated in the manuscript in ll.337-341.

24. - l. 277 (we did not investigate precip directly): I think you need to be upfront about this caveat and discuss it in the methods section
Thank you for your nice suggestion, now this discussion has been added in the Methods section (see ll.155-159).

25. - l. 304 ("reduced tropical influence on the event"): As mentioned before, it would be good to have some further analysis on this.

Please see our response above.