Response to Editor

Dear authors,

Thank you for submitting the revised version of your paper to GMD.

As you can see, one of the reviewers has provided a thorough and constructive review of your revised manuscript. I agree with the reviewer that it is necessary to evaluate the short-term skill of the model, rather than relying on climatological precipitation, and also use a statistically independent dataset. Both issues need to be addressed.

Regarding code and data availability, I also agree that although GPM and ERA5 can largely be considered standard datasets in atmospheric sciences, it would be appropriate to make the specific dataset you used available via a Zenodo archive in order to make the results of the paper fully reproducible.

Kind regards,

Emmanouil Flaounas

Dear Editor and Reviewers,

Thank you for your time, review, and consideration of our manuscript. We have revised the paper in response to your comments. In particular, we re-evaluated the offline training performance using a fully independent GSRM dataset that is outside the 80-day simulation samples. We also added an assessment of the short-term forecast skill of GRIST-MPS versus GRIST-CPS, which is provided in the Supplementary Materials. The specific dataset we used has already been stored at Zenodo, and is now explicitly stated in the code and data availability.

All other comments have been addressed point by point in the response.

All the best

Yi Zhang (on behalf of the authors)

Response to Reviewers' Comments

Response to Reviewer #1:

We thank this reviewer for the insightful comments and detailed suggestions on how to improve the manuscript. The manuscript has been further improved based on your comments. Below is the item-by-item reply to your questions and suggestions, the texts with normal font are your original comments, the texts with blue font are our responses and the texts with italics are the revised content of the manuscript.

Line comments:

L118: This split does not yield an independent validation dataset. Neighboring time points in your dataset are incredibly correlated, especially because they are derived using linear interpolation from a coarser time grid. The linear interpolation aspect means you might as well have used the training dataset itself as the validation dataset. This is not a reliable assessment of out-of-sample performance. Assuming Figure 2 is based on validation data, it must be reevaluated using a statistically independent validation dataset, as this is your main measure of process skill.

Reply: We appreciate the reviewer's suggestion. Accordingly, we introduced a completely independent 10-day GSRM dataset as a new validation set to re-evaluate the model's out-of-sample performance. We have updated Figure 2 in the main text accordingly.

As also shown in Figure A1, the overall patterns remain consistent between the two validation sets. While the R^2 values from the new validation set (right panel) are slightly lower in certain regions compared with those from the previous validation set (left panel). This indicates that the model retains a reasonable degree of reliability on out-of-sample data.

Corresponding revisions have also been incorporated into the manuscript. (L119-122).

"Moreover, we have further conducted a fully independent 10-day GSRM experiment (date: 14–23th, July 2008) for final validation to assess the model's out-of-sample performance (see Fig. 2 for details)."

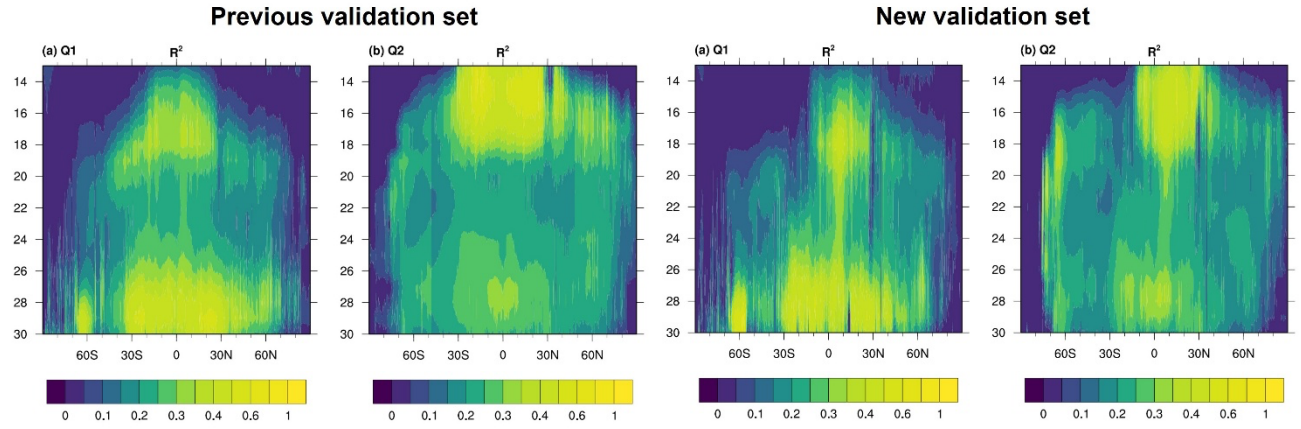

Figure A1. Offline skill of the coefficient of determination (R^2) for Q_1 and Q_2, as functions of latitude and model level. Left panel: the previous Figure 2, Right panel: as in Figure 2 but for new independent 10-day GSRM validation set.

L166: Remove "per epoch" from "weight decay of 10^-6 per epoch". Adam weight decay is applied on each gradient update, not just once per epoch.

Reply: Thank you for your comments. Done.

L167: Please define "superior performance". What metric on what data?

Reply: We have now added the performance and data descriptions to the revised manuscript (L167-168).

"The mean absolute error (MAE) loss was selected over the mean squared error (MSE) loss as the loss function, as it demonstrated superior performance with lower biases on validation sets during initial training phases."

L184: What "offline performance benchmarks" - what metric(s) on what data?

Reply: Offline performance benchmarks refer to the metrics of MSE < 1×10^{-4} and $R^2 > 0.3$ achieved on the validation sets as described above. This content has also been incorporated into the corresponding paragraph. (L184-186)

"The final selection of our optimal MPS is based on a dual evaluation: satisfying offline performance benchmarks (MSE < 1×10^{-4} and $R^2 > 0.3$ on validation sets) and demonstrating stability in online integration, which must maintain stable online integration for more than 3 months."

L186: "can integrate stable" -> "can integrate stably"

Reply: We have revised it to "demonstrated stable integration".

L186: Please define "better performance", what metric on what data? Is this on validation data, or is this a metric applied to the online simulation? To 6 years of simulation or a longer/shorter period?

Reply: "Better performance" refers to the ability to perform stable integration for more than 6 years, with better RMSE values for online variables such as precipitation. Relevant details have been added in the manuscript (L187-189).

"*Among the eight NNs, two demonstrated stable integration capabilities for over six years. We selected the better-performing model between them as the optimal MPS, based on its improved RMSE values for online variables such as precipitation (see Table 1).*"

L237-244: Please comment in the manuscript on whether the model is unstable or simply performs poorly when the near-surface or top-of-atmosphere levels use MPS instead of CPS.

Reply: The model is unstable when the near-surface or top-of-atmosphere levels use MPS. The relevant descriptions have been added in the manuscript. (L241-242)

"*This selective preservation, validated through sensitivity experiments, serves as a stability-enhancing mechanism, the present MPS without this configuration will crash after running for a few days.*".

L267-273, Figure 4: In several areas away from the pacific ITCZ region, GRIST-MPS results look clearly worse than the GRIST-CPS values. This is borne out in the pattern correlation coefficients, which indicate 26% and 12% unexplained variance of the pattern respectively - more than double in the MPS-based model. Specifically, there are large dry biases in the northern and southern extratropical ocean winter regions, and a large moist bias over Africa in both seasons. Provided this is an accurate assessment, please ensure these are remarked on, or add to the paragraph on lines L288-290.

Reply: Thank you for this remark. The content regarding the limitations of GRIST-MPS has now been incorporated into the manuscript. (L287-289).

"Meanwhile, some remaining biases persist in GRIST-MPS: compared to GRIST-CPS, its global spatial pattern correlation coefficient is marginally lower (PCC: MPS: 0.86 vs.CPS: 0.94) on Summer. For it exhibits substantial dry biases over the northern and southern subtropical oceans in winter, alongside a consistent moist bias over Africa in both seasons."

You may want to separately remark on the pattern correlation coefficient confined to the pacific ITCZ region, which I would expect to be better, reflective of the improved ITCZ strength and width.

Reply: We calculated the pattern correlation coefficients for the Pacific ITCZ region, the results show that CPS yields better performance in summer (MPS:0.89 vs. CPS: 0.93) , whereas MPS exhibits a more robust simulation capacity in winter (MPS:0.95, CPS:0.94). We attribute this to the fact that although both the strength and width have been improved in both seasons, there remain certain biases in the position of the ITCZ compared with GPM during summer. Such biases give rise to a double penalty issue, which ultimately results in its correlation coefficients being lower than those of CPS. For this reason, the correlation coefficients for the Pacific ITCZ region are not presented in this section; instead, the ITCZ and spatial correlation are introduced separately.

L268-269: Wodzicki and Rapp 2016 do not define ITCZ strength in this way. In section 2.3 they specify that precipitation intensity should be used to measure ITCZ strength, but they take this as a monthly mean value across grid cells in a specific ITCZ region. Please use established measures for ITCZ strength and width.

Reply: The strength and width of the ITCZ have been recalculated using well-defined methods, with appropriate citations provided for each (L271-276).

"*Following established conventions, we quantify the width of the ITCZ as the latitudinal distance between its northern and southern boundaries, applying a 5-point smoother to the data prior to calculation (Wodzicki and Rapp, 2016). And the northern (southern) boundary is identified by moving equatorward from higher latitudes and locating the first grid cell where the precipitation in the adjacent grid cell to the north (south) falls below 2.5 mm/day. The strength of the ITCZ is defined as the area-weighted mean precipitation within the region bounded by these northern and southern boundaries (Wang et al., 2023).*"

L270: The cited Byrne et al. 2018 paper is a review paper which does not define ITCZ width. They reference one paper which defines ITCZ extent using a precipitation minimum - Wodzicki and Rapp 2016 (https://doi.org/10.1002/2015JD024458), which you cited for precipitation strength. This paper reviewed a range of rainfall thresholds (1, 2.5, and 5.0 mm/day) and found 2.5 mm/day to be most consistent across datasets, and used a 5-point smoother prior to taking the threshold. Please use this methodology for measures of ITCZ width and cite this paper, or cite another paper supporting a different methodology.

Reply: We appreciate your suggestion. We have now applied the method you suggested, and provided a clear description of it in the paper, as detailed in the previous response.

Summary and outlook:

It is hard for me to agree with publishing the characterization of model skill given in this summary. While it is true that the model more accurately represents the Pacific ITCZ rainband (thank you for improvements to the manuscript which quantify this), significant precipitation biases are produced in most other regions with strong precipitation features, worse than the CPS-based model. This is borne out in the pattern correlation coefficients. These should be given proportionate weight in the assessment. The relevance of this work for future ML schemes could be discussed in more detail.

Reply: Thank you for your comments. The limitations of GRIST-MPS have been clearly indicated in the manuscript(L287-289), and the discussion on the relevance of this study for future ML schemes has also been reorganized accordingly(L392-410).

"Meanwhile, some remaining biases persist in GRIST-MPS: compared to GRIST-CPS, its global spatial pattern correlation coefficient is marginally lower (PCC: MPS: 0.86 vs.CPS: 0.94). For it exhibits substantial dry biases over the northern and southern subtropical oceans in winter, alongside a consistent moist bias over Africa in both seasons."

*"**Prospective Relevance.** As a proof-of-concept study, this work provides useful reference for future ML-based efforts by demonstrating that a column-based ML physics module, trained on GSRM multiscale modeling data over a limited time window, can produce realistic online free-running climate simulations and generalize to unseen periods. Beyond the choice of ML architecture, several strategies proved essential. First, spatial random sampling mitigated spatial overfitting arising from an imbalanced data distribution, improving both accuracy and stability. Second, interpolating the data from 1-hour to 20-minute resolution tripled the effective sample size, imparted a regularization effect, and aligned the training data with the model time step, further improving online performance. Notably, the model still requires a small number of near-surface CPS tendencies (primarily for the boundary layer) to maintain online stability. This likely reflects a limitation of the coarse-grained dataset, which supports only a coarse vertical resolution in the lower troposphere (spacing >200 m below 850 hPa). In addition, vertical eddy transport associated with boundary-layer turbulence may require higher-resolution training data to be represented effectively."*

"*An important implication is that representing physics tendencies as residual terms of grid-scale variables is a promising route for diagnosing training targets for ML-based physics. This approach offers several potential advantages that merit further exploration. For example, a common high-resolution dataset can be incrementally coarse-grained to multiple coarser resolutions, enabling datasets at different resolutions to share consistent large-scale information. Moreover, the large-scale states can be further constrained (e.g., via short-period simulations and/or nudging), which may improve the quality of the diagnosed tendencies, and/or generate additional constraint-related tendencies that can be learned separately. However, unlike physics tendencies directly extracted from a host model, this reconstructed-tendency approach generally requires dedicated training strategies and careful coupling procedures to ensure stable and accurate online integration.*"

Code and data availability: The editor should comment on this, but is it sufficiently reproducible to provide a simple web link to the top level page for GPM and ERA5 data? Websites are subject to change, the process for downloading this data is quite involved, and at least in the case of ERA5 there are several data products to choose from.

Reply: The specific ERA5 and GPM datasets used in this study are also available in the linked code repository, within the compressed archive **input_plot.tar.gz**. This has been now explicitly mentioned in the data availability.

"The specific GPM and ERA5 datasets used in this work were archived in the file input_plot.tar.gz on the Zenodo repository provided."

Overall review:

It looks like a wash whether this model outperforms the CPS-based model on climate skill - arguably the precipitation climatology is worse. I do not think this on its own is a reason to reject this manuscript. However, it is important that the manuscript does not give a false sense of the climate performance.

Reply: Thank you for your valuable comments. In the original manuscript, we presented results primarily emphasizing where the GRIST-MPS model outperformed the GRIST-CPS model. As you noted, this argument is not fully conclusive based on the present simulation, but it does indicate a

positive signal of learning physics from GSRMs.

In the revised version, we have accordingly strengthened the analysis of precipitation climatology and, more importantly, elaborated on the limitations of the GRIST-MPS model to provide a more balanced and objective evaluation.

"Meanwhile, some remaining biases persist in GRIST-MPS: Compared to GRIST-CPS, its global spatial pattern correlation coefficient is marginally lower (PCC: MPS: 0.86 vs.CPS: 0.94). For it exhibits substantial dry biases over the northern and southern subtropical oceans in winter, alongside a consistent moist bias over Africa in both seasons."

I do think it's an achievement to produce ML-based column physics trained on only 80 days of data which does anything remotely reasonable when run online, and has any accuracy on out-of-sample process representations. It is likely the use of CPS near the surface and top of atmosphere is crucial for this, as well as the use of interpolation to augment the training dataset or regularize the model. If this is the main achievement of the paper, it is important to demonstrate that these are key components by commenting on the performance or stability when these features are omitted.

Reply: We appreciate your recognition. The experiments you mentioned have been conducted and are summarized in Table 3, with the exception of the test using the near-surface MPS, which led to rapid model instability.

As shown in Table 3, both random sampling and temporal interpolation contribute significantly to model performance: random sampling effectively improves stability, while temporal interpolation primarily enhances accuracy, and also suggests better generalization ability in models trained with this approach. Corresponding discussions of these findings are included in the concluding section of the paper.

"***Prospective Relevance.*** *As a proof-of-concept study, this work provides useful reference for future ML-based efforts by demonstrating that a column-based ML physics module, trained on GSRM multiscale modeling data over a limited time window, can produce realistic online free-running climate simulations and generalize to unseen periods. Beyond the choice of ML architecture, several strategies proved essential. First, spatial random sampling mitigated spatial overfitting arising from an imbalanced data distribution, improving both accuracy and stability. Second, interpolating the data from 1-hour to 20-minute resolution tripled the effective sample size, imparted a regularization effect, and aligned the training data with the model time step, further improving online performance. Notably, the model still requires a small number of near-surface CPS tendencies (primarily for the boundary layer) to maintain online stability. This likely reflects a limitation of the coarse-grained dataset, which supports only a coarse vertical resolution in the lower troposphere (spacing >200 m below 850 hPa). In addition, vertical eddy transport associated with boundary-layer turbulence may require higher-resolution training data to be represented effectively.*"

The use of highly correlated validation data is still a major issue in the manuscript which must be addressed. Figure 2 must be produced using samples which are not correlated to the training dataset. If this figure does not show skill when evaluated on independent examples (which is a possible result), it would not be reasonable to publish this manuscript. Ideally uncorrelated out-of-sample validation would be completed throughout training process, to assess whether the model is overfitting and at what point during training. This is not being done with the described train validation split.

Reply: As replied in our earlier response, we have re-evaluated Figure 2 using a fully independent

GSRM simulation dataset, out-of-the-80day-samples. The results (Figure A1) show that the R^2 values for Q_1 and Q_2 in this independent dataset closely align with our previous findings in both pattern and magnitude, confirming the good generalization capacity of the trained model.

We fully agree that incorporating uncorrelated out-of-sample validation throughout training is essential for quantifying overfitting and identifying the optimal stopping point. Given the time constraints and the model's satisfactory performance in the independent validation described above, we did not retrain the ML model for this study. Nevertheless, the proposed training-to-integration workflow is readily extensible and can be straightforwardly applied in future work with systematic out-of-sample validation.

It is also concerning to have the manuscript state it uses previously-established measures for ITCZ strength and width, but then not use the measures described in the cited papers. This type of error is easy to miss in review as it is not the reviewer's responsibility to read cited papers. It is necessary to use established measures for complex measures like this to avoid p-hacking (which I don't think is the case here, the ITCZ representation does look better). Perhaps I am not understanding the measures in the cited papers, but this should be explained.

Reply: Thank you for your comment. We acknowledge that the definitions of ITCZ intensity and width in the original manuscript were not sufficiently clear. In the revised version, we adopt more explicit, literature-based definitions with appropriate citations to ensure clarity for you and other readers. Results using the updated definitions are consistent with our original findings and do not affect the study's conclusions. For further details, please refer to our earlier response.