Review of Global Climate Modeling with Improved Precipitation Characteristics by Learning Physics (GRIST-MPS v1.0) from Global Storm-Resolving Modeling

The model configuration proposed by the authors is sound and worthwhile to pursue, with a high quality training data source. Specifically, it would be a highly impactful and somewhat surprising result to demonstrate that MPS can be trained to produce a skillful online parameterization using only 80 years of GSRM training data, especially one using deep learning approaches which require many independent samples to train without overfitting.

However, there are presently critical issues in the evaluation of the methodology that make it unclear the methodology has produced a skillful model, or the extent to which any model skill may be a consequence of p-hacking (due to a lack of detail about model selection strategies).

In my review I have provided specific examples of how the evaluation could be improved to demonstrate the model has skill. The authors should also make it clearer that the trained model was not selected based on the measures shown by explaining exactly how it was selected. I strongly suggest the authors use a validation dataset from the GSRM congruent with the training data to verify the model is not overfit.

Abstract:

- The abstract is light on some specific details which would be valuable to a reader in understanding the contribution of the paper. While these may not be strictly required, adding them would significantly improve communication of the results.
- Should mention the resolution of the GCM/ML, and that the GSRM simulation data is coarsened (coarsening is itself a difficult problem you are tackling here).
- (Optional) It is also worth highlighting the use of full topography.
- (Optional) What is the GCM used?
- (Optional) Says comparison is being done against observations, it would be worth spelling out precipitation observations and historical reanalysis/ERA5. Currently reanalysis comparison is not mentioned in the abstract.

Overall comments:

- To evaluate the quality of the learning, the method output must be compared against the GSRM configuration of the model, which is congruent with the training procedure. Observations and long integrations may also be compared against, but it needs to be clear the extent to which observations are being matched by properly learning the training data, as opposed to just having less variability. This can only be done to the extent that the match of the GSRM with observations or reanalysis is known/presented. For example, does the GSRM have reduced precipitation compared to the CPS configuration? Does the mean precipitation of the long-term MPS configuration agree with the precipitation present in the GSRM training set? One would imagine it is not biased in the first few timesteps - does the precipitation produced by the MPS configuration stay consistent throughout the run, or does it reduce as the simulation leaves the initial condition?

- Consider including a baseline similar to MPS but where the ML-based predictions of Q1 and Q2 are uniformly zero. At the least, the MPS model should better match the coarsened GSRM data than this baseline, and this comparison shows whether the MPS has any positive skill when integrated online (whether or not it outperforms CPS).
- The word "validation" only appears once in the manuscript, and it is not clear what validation dataset if any are used to measure overfitting and skill during training and hyperparameter optimization. On L91-92, it is implied that GSRM-style data is only used for training, and not for evaluation, making it hard to understand how out-of-sample performance on the loss could be evaluated. I am well aware of the high cost of GSRM output, but at least a few days of independent simulation data should be used for this purpose. Especially given the possibility of overfitting on a dataset only 80 days long, it is crucial to include validation metrics during training.
- The implication of the lack of validation dataset is that the authors tuned the model on the measures being presented in the paper. If this is true, it significantly increases the potential any model skill is due to random chance, especially given that the final model was selected by testing an unnamed number of models.

Line:

- A few lines talk about scale-invariance, though none is shown here. L45: "Ideally, such models would not only perform robustly at a specific resolution but also enable a smooth transition across multiple meteorologically significant scales, from the typical GCM resolution (100 km) to the GSRM resolution (1 km)" and L119 "Although the present study coarse-grains GSRM data to a fixed resolution, the residual method allows efficient transitions 120 from arbitrarily high-resolution models to GCM target scales, thereby enabling the MPS to be inherently scale-aware." It is not clear to me here or from referring to Zhang and Chen 2016 how this is the case. For instance, a model which learns residuals at a given resolution cannot be applied zero-shot at other resolutions, any more than a model that learns the full field. The model does not appear to be scale-aware as mentioned on L120. I would suggest removing these references to smooth scale transitions, as they are not important to the primary contribution of this paper. If they are kept, they should be supported at least with a theoretical basis, if not experimental results. Note that the results here do successfully lead a coarser model to behave more like a finer model, but this is not what is generally meant by scale-awareness or by a smooth model transition across scales.

- There are some conflicting statements about the configuration of the GCM - it's an important detail whether it's using explicit convection or a cumulus scheme. On L93-94 it says the cumulus scheme is disabled, but on L94 the GSRM is compared against the GCM with a conventional cumulus parameterization. On L103 it mentions the cumulus parameterization being turned on. Any results of MPS should be compared in skill against the optimal GCM configuration for the base model, in this case it seems one with the conventional cumulus parameterization turned on, even if MPS is hooked into a version of the GCM without cumulus parameterization. I think that's what you're doing, but perhaps L93-94 should be modified.
- L117-118: How are Q1 and Q2 computed numerically? For example, are the terms computed by taking gradients of the coarsened fields according to a given numerical scheme (if so, something like "where gradients are computed using [...]"), or by taking the tendency of the dynamical core in coarse mode?
- L142: Please clarify whether model outputs are referring to GSRM values or ML outputs.
- "column" only appears as of line 146, and I only realized this is a column model by line 156. The model being column-local is noteworthy and should be mentioned much earlier.
- L152: What optimizer is used?
- L152: How was model skill evaluated when performing hyperparameter optimization?
- L153: The stated decay rate of 1e-6 per epoch is either incredibly large (if directly multiplicative, giving a final epoch LR of (3e-4)*(1e-600), or incredibly small (if the multiplier is (1 - rate), giving a final epoch LR of 2.9997e-4). Is this accurate?
- L153-154: it is conventional to say you are using a MAE loss function, and how the values are normalized when computing the loss (in this case, it would appear to be using the min-max scaling mentioned elsewhere). It could be worth remarking on why you chose MAE loss over MSE, if there is a reason.

- L160: Please mention the batch size here alongside the number of samples per epoch.
- L162-172: You mention a process of training many models and evaluating for the best candidates. Please specify how many models were trained for this selection process.
- L171: Please specify exactly what is meant by “demonstrating stability in online integration”. For example, was there a pass/fail measure of stability? Does the paper present the first trained model for which this stability measure was passed?
- L173-183: No edit needed here, but this result is quite curious and I am skeptical of the interpretation, since the linear interpolation doesn’t grant any true additional temporal resolution, and emulator models successfully operate at 6-hour timesteps. It may be that the model behaves better online when it must make 3 of roughly the same (residual) prediction, and its errors get averaged or the intermediate values serve as a type of data augmentation. You might consider for your own purposes training this model with a 1-hour timestep but linearly interpolating the training data from 3-hour intervals down to 1-hour, and see if you get a similarly stable model. This would prove out that it is in fact the 20-minute temporal resolution which is key, and not the structure of linearly-interpolated data.
- L186-188: It is not right to say the effective number of training samples is tripled, given these are not independent samples but rather completely linearly dependent on existing training samples. Rather, this is similar to other regularization strategies used for improving stability, e.g. https://arxiv.org/abs/1912.02781 or Bishop (1995) “Training with Noise is Equivalent to Tikhonov Regularization”
- L215-216: It is somewhat problematic to compute precipitation in this way, as the value is really precipitation minus evaporation. This significantly affects the headline result of the paper (improved preciptitation skill through lower RMSE). It would be appropriate to compare against precipitation minus evaporation for the CPS configuration, and to appropriately describe these values in later sections. For example, this significantly affects the interpretation of Figure 7.
- L223-225: Please quantify the reduction in computational cost of GRIST-MPS (using CPS for the layers described) compared with GRIST-CPS.
- L228: What years are used for the AMIP-style simulations? Are these related to the years used for the GSRM training simulations?
- L243-247: Please comment on the relative strength of the ITCZ between the models, and quantify the underestimation and overestimation of precipitation in the southern oceans.

- L249-251: Can you give evidence this reduced RMSE is due to improved placement of the ITCZ and not reduced strength, for example by using a measure that accounts for the variance of the model?
- Figure 4: GRIST-MPS seems to have significantly reduced mean precipitation, but it is hard to tell from color maps alone. Please include a measure of mean bias, as well as a measure of skill which does not improve for models with less variability, such as the Pearson correlation coefficient. This would make it clear the MPS model is outperforming the CPS model.
- Figure 5: The rainbow color bar is not doing you any favors here, making it look like you have a drastic reduction in precipitation. Using a more linear color bar would give a more accurate representation of the data.
- L259-260: Does this really show superior constraint of the ITCZ width with MPS? The green contour lines appear similar to CPS, and past that the ITCZ appears weakened more than narrowed. The lack of a maximum around -15S Jan-March for MPS is also notable, weakened variabillity in this region is also apparent in Figure 4.
- Figure 7: Please include the coarsened GRIST data used as target for the MPS model. It is not clear whether the MPS is doing “better” because it is accurately representing reduced precipitation as seen in the GSRM, or because it has weakened variability in general compared to CPS.
- L309-310: Please quantify improved spatial pattern alignment. RMSE is not sufficient for this.