Review of egusphere-2025-2790

Overall review follows after line comments.

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 inteprolation 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 re-evaluated using a statistically independent validation dataset, as this is your main measure of process skill.

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.

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

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

L186: “can integrate stable” -> “can integrate stably”

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?

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.

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 L282-284.

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.

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 gridcells in a specific ITCZ region. Please use established measures for ITCZ strength and width.

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.

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.

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.

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.

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.

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.

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.