Second Response to Reviewers: ”Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators”

April 3, 2025

Overview

We sincerely appreciate the reviewers’ thoughtful feedback and thorough evaluation of our paper. In our prior response (https://egusphere.copernicus.org/preprints/2024/egusphere-2024-2420/egusphere-2024-2420-AC2-supplement.pdf), we included our outlined plan for changing the manuscript (e.g. changing text, adding new figures, updating existing figures). We have now submitted our revised manuscript. In this document, we include pointers to the updated manuscripts that resolve the reviewers’ major comments: we do not include all the reviewers’ comments here. For a complete line-by-line discussion of all the reviewers’ comments, please see our prior response at the link above. Our manuscript changes are in green, and the reviewer comments are in black.

Changes to the Manuscript In Response to Reviewer #1

Page 2: The ML model has “orders-of-magnitudes” lower computational cost. Is this really true? More than a factor of 10? This could only be possible if the IO cost (that will stay the same) is considered to be of less than 10% of the overall cost (also see comment for Part 2). And what is the “cost”? Time, energy, or hardware purchase?

See the Part II changes, where we add a sentence on the computational efficiency of ML that makes it feasible to generate 256 members in one minute in parallel on 256 GPUs.

P6, paragraph starting with “We choose SFNO…”: I find this part difficult to follow. It would be good to remind the reader about the architecture of the SFNO.

We have now included a description of the SFNO architecture in section 2.1

What exactly do you mean by downscaling and scale factor here (I think I know, but only since I know the previous papers)? I do not understand why a lower scale factor would lead to a larger ensemble spread.

We added this to the architecture description in Section 2.1. Also, on page 7, we have included the following text:

The scale factor controls the level of spectral downsampling of the input field. With more aggressive downsampling, SFNO internally represents the input atmospheric state with reduced resolution. We speculate *that this may reduce the effective resolution of the predictions. With a reduced effective resolution, small-scale perturbations would not grow and propagate upscale. Instead, they would be blurred out, and they would not result in increased spread among ensemble members.*

Figure 10: Is the control member equivalent to a normal ensemble member, or are there small differences (as in IFS)? Can you also plot 0h?

See Figure D5. Also, see associated text:

At a lead time of 360 hours, the perturbed members maintain similar spectra as the control member (Figure 10), and at the initial time, they have similar spectral characteristics as the unperturbed ERA5 initial condition (Figure D5).

Figure 2b: For what timestep are the spectra calculated?
They are calculated at 360 hours. We have added this information to the figure caption for Figure 2b.

Changes to the Manuscript In Response to Reviewer #2

Despite these promising results, several aspects warrant further research. The authors mainly focused on 2m temperature, especially heat extreme from model configuration to diagnostics, and given that the authors deliberately included 2m dewpoint temperature as a model input variable, incorporating predictions of derived heat extreme indices would provide valuable insights into the model's capabilities.

We have included diagnostics on the heat index in Appendix D, Figure D2

Furthermore, I recommend evaluating a broader range of LLHIs to strengthen the reliability of the approach. I think the authors can incorporate cold extremes along with heat extremes. What about wind extremes, which are in prediction variables?

We have included diagnostics on the 10m wind speed and cold extremes in Appendix D, Figure D3 and Figure D4. In the appendix and in the main text, we state that the model performs well on these other variables at 48 and 96 hours. At 240 hours, the model's reliability degrades for probabilities greater than approximately 50%. This is the subject for further research. In addition to appendix D, we highlight this in the main text also on page 23:

We visualize the reliability diagrams for other lead times (Supplemental Figure D1) and variables. We show that SFNO-BVMC also performs reliably when forecasting the heat index at lead times of 48, 96, 120, and 240 hours. For 10m wind speed and cold extremes, SFNO-BVMC matches the performance of the IFS ensemble (Figure D2 and Figure D3). However, we also show that at 240 hour lead times, the model is not reliable when it confidently (greater than 50% chance) forecasts wind extremes or cold temperature extremes (see Appendix D and Figure D4 for more discussion). This is an area for future model development.

As well as various extreme events, actual forecasts would be helpful to recognize the usefulness of the model. Diagnostics with real-event prediction would be more persuasive. For example, t2m ensemble time series at a certain grid point, trajectories of each ensemble for each variable, and the difference between IFS could strengthen the model's credibility.

We have included a real-event prediction demo in Appendix A.

Major Comments 1. (p.4) “Existing work has shown that simple Gaussian perturbations do not yield a sufficiently dispersive ensemble. (Scher and Messori, 2021; Bülte et al., 2024): the ensemble spread from these perturbations is too small.” If so, you can still adopt singular vectors or other methods to reflect initial condition uncertainty. Are bred vectors superior to other approaches? Are they the cheapest way other than simple Gaussian perturbations?

Thank you very much for raising this helpful suggestion. We have added this to the discussion section, page 26:

Understanding how ML models respond to perturbations is an important research frontier. In particular, future work is necessary to compare the computational cost and skill of different initial condition perturbation methods, in tandem with model perturbations. We find that bred vectors are a computationally inexpensive way to achieve reasonable spread-error ratios and to generate an arbitrarily large ensemble. Further refinement of initial condition perturbation techniques is needed to improve forecast performance.

We continue the discussion by comparing bred vectors to other perturbation methods.

4. (p.10) Figure 3. The ensemble spread from different numbers of checkpoints. : Model configuration also focused on 2m temperature. Do we need to change the number of checkpoints if we want to forecast wind extremes? Do we need to change it every time for different variables? Selecting the number of checkpoints based on the comparison among multiple variables would be a more optimal choice.

We have added 2 additional variables to our analysis to Figure 3.

5. (p.17) “On the second criterion, crucially, their spectra remain constant through the 360-hour rollout (Figure 10 and Figure 11).”

: Degradation of power in short wavelengths occurs in a lot of DLWPs. Then are all DLWP models' degradation because of autoregressive fine-tuning? This seems like a crucial problem to just hypothesize the cause. I think it would be beneficial for readers to pinpoint the cause.

6. (p.17) “While the control and perturbed spectra remain constant through the rollout, the SFNO-BVMC ensemble mean does increasingly blur with lead time. Figure 12 shows that the ensemble means of SFNO-BVMC and IFS ENS similarly degrade in power after 24 hours, 120 hours, and 240 hours.”

: In the first paragraph of section 3.2 Spectral Diagnostics, the authors elaborate that power decay is one of the symptoms of blurriness, but this sentence seems like presuming those two are equivalent. section 3.2 needs to be more clear. What is the relationship between spectra and blurriness in general and what did SFNO find? Why is SFNO-BVMC different from other DLWPs with respect to the power spectrum?

We have completely rewritten ”Section 3.2 Spectral Diagnostics” on page 17 to address these questions.

7. (p.19) “This is necessary but as yet insufficient validation for our main scientific interest in LLHIs.” : I expect more analysis of LLHIs such as case studies that occurred during recent years, even though the authors agreed with the lack of validation. It would provide a more robust evaluation and help illustrate the model's practical value.

We have included a case study in appendix A.

Minor Comments

1. (p.12) “First, they contain a land-sea contrast for surface fields such as 10m wind speed and 2m

temperature. For these surface fields, perturbations have distinct amplitudes and spatial scales over the land and ocean.”: It’s a bit difficult for me to discriminate the difference. Could you show the amplitude in another way?

We have made the following change:

For these surface fields, the perturbations have distinct amplitudes over the land and ocean. In this example, the 2m temperature perturbation has an amplitude of 0.56 K over land and 0.27 K over the ocean, and the 10m wind speed perturbation has an amplitude of 0.45 m/s over land and 0.66 m/s over the ocean.

References