Response to Reviewers: "Huge Ensembles Part II: Properties of a Huge Ensemble Designed with Spherical Fourier Neural Operators"

February 17, 2025

Overview

We sincerely thank the reviewers for their constructive comments and review of our paper. These comments will substantively improve our manuscript. We have included responses to the reviewers' comments below, with the reviewer comments in black text and our response in green text.

In this document, we will detail our planned revisions. Some of the reviewer comments require re-analyzing our ensemble simulation with additional variables, such as 10m north-south wind, 10m east-west wind, heat indices, and others. We request a short period of time due to the computational and data requirements of these comments. The ensemble simulations use O(1) TB per variable for a 58-member ensemble, and O(100) TB per variable for the huge ensemble. Upon completion of the analysis that the reviewers suggest, we will submit a revised version of the manuscript in mid-March 2025.

For public reference, this is the second of a two-part manuscript on huge ensembles. We refer to part I as HENS Part I [Mahesh et al., 2024a], and we refer to part II has HENS Part II [Mahesh et al., 2024b].

Comments from Reviewer #1

Part 1 and 2 are both interesting papers that document the development and use of a machine learned ensemble weather forecast model with an enormous number of ensemble members. The papers fit well into GMD, but I think that they should be revised following the comments below. The paper presents very interesting and useful information about how the huge ensemble is generated and how much effort it requires to run such a model.

Thank you for this overview of our paper.

However, the evaluation of the usefulness of a huge ensemble is rather weak as it is presenting the "easy" task of Gaussian predictions but avoids diagnostics that evaluate the "hard" tasks for a huge ensemble that could actually show the real usefulness. I am left a bit puzzled after reading the paper how the huge ensembles could actually become useful. I doubt that our current 50-member ensembles would greatly benefit from more ensemble members if we assume that we predict Gaussian distributions.

Thank you. We wish to clarify two things:

1. We do not assume that the ensemble distribution is Gaussian. This is an emergent property of the ensemble quantity we study in Figure 2 and Figure 4: the global land mean for each variable. The Gaussianity could arise from taking a large spatial mean over all land grid cells over the globe. We state,

For all variables, the HENS gain closely follows the theoretical Gaussian gain. This result is not completely surprising: averaging over a large number of grid cells implies that a Central Limit Theorem should apply (even though the grid box values are neither independent nor identically distributed), wherein the global land averages behave similarly to a Gaussian random variable. In Section 3.1, we state why we choose to study this particular quantity:

At a given time, there will likely be extreme conditions occurring somewhere on Earth, simply due to the spatial variation of weather. In our calculation of information gain, we do not consider the spatial distribution of extremes, which varies significantly within each ensemble member. Instead, we wish to assess the distribution across ensemble members. Therefore, we calculate the information gain on the global land mean values of each ensemble member. This allows us to assess the ensemble members in aggregate and how far each ensemble member is from the ensemble mean.

2. For all other analysis, other than Figures 2 and Figures 4, we analyze statistics at the grid cell level. In particular, Figure 3 is a version of Figure 2 without making the Gaussian assumption or taking a land-mean. We make no assumption about the distribution of the quantity of interest, neither in the ensemble forecast nor in the verification ERA5 dataset.

We have EMOS to improve predictions for 50-member ensembles, so no need for huge ensembles. How does the information gain of 4 compare against the IFS ensemble with EMOS?

For our revised manuscript, we are working to obtain enough of an IFS model climatology to calculate this quantity. However, for SFNO-BVMC, we note that gain is convincingly a function of ensemble size, both with and without the Gaussian distribution at play, (Figure 2 and Figure 3).

It seems to be of less relevance to have a prediction of the uncertainty range of the probability for an extreme prediction.

We respectfully wish to highlight this as an important advantage of HENS. Taking our example heatwave in Shreveport, Louisiana, USA, a 58-member ensemble predicted that a heatwave could occur with 18% probability, but the 95th percentile confidence interval of this event was large, from 8.7% to 28% (Figure 9 and section 4.2). In the extreme event forecast, there is an uncertainty introduced by limited sampling: in the space of all possible ensemble members, which 58 members were selected? HENS reduces this sampling uncertainty significantly. For Shreveport, the confidence interval became narrower, ranging from 17.1% to 18.9%; see Figure 10 for the results of this analysis across the entire period.

HENS and the 58-member ensemble are almost equally reliable (Figure C1), but HENS has narrower confidence intervals. In this way, HENS is a more useful forecast. We propose that reducing the error bounds on extreme event forecasts is a highly valuable benefit of huge ensembles.

I also do not think that an ensemble range that predicts temperatures between 295 and 320K will be of any assistance for a decision maker (as seen in Figure 5). To have a couple of members from a 1000-member ensemble close to the truth will not trigger any decisions for a forecast. The same is true for the outcome-weighted CRPS discussion. If we assume that the distributions of variables that are of interest are non-Gaussian, in particular for extremes, the huge ensembles may be extremely useful to sample the tails of the distribution.

We agree: if a couple of members from a huge ensemble are close to the truth, that will not trigger decisions for a forecast because we do not know which members those will be ahead of time. For an operational decision-maker, the benefit of a huge ensemble is that it reduces the error bounds on the probability of extreme (see point above), and that it has a reduced outcome-weighted CRPS (owCRPS). owCRPS directly measures how well the forecast distribution resolves events that are above the extreme threshold, and it does not rely on the Gaussian assumption. It measures the performance of the ensemble at the tail of the verification distribution (ERA5).

While a huge ensemble has a large temperature range by design, an operational decision-maker can use the probabilities associated with these temperatures to inform their decisions. In particular, they can use the huge ensemble if they are particularly concerned about low-likelihood events or to reduce the sampling uncertainty associated with extreme events. In the paragraph above, we discuss HENS's performance at the tail of the verification distribution (ERA5). In addition to this, HENS directly samples the tail of the forecast distribution with higher fidelity (e.g. Figure 9). This allows HENS to sample events that are low-likelihood in the forecast. These low-likelihood events do occur: 3% of events in summer 2023 were low-likelihood enough that they were completely out of the bounds of the IFS 10-day forecast (Figure 11b). HENS can sample these events (Figure 11a) without a degradation in CRPS or reliability. This means that HENS offers a way to directly simulate events that are out of the bounds of IFS. For these events, it provides a better estimate of low-likelihood events that exist at the tail of the forecast distribution. By simulating these events, HENS also offers a dataset with a large sample size to study the dynamics of these low-likelihood events (at the tail of the forecast distribution) post-hoc.

We also emphasize that HENS can be used beyond the context of operational meteorological decision-making. It can be used to study the drivers and statistics of extreme events in a retrospective hindcast mode. For this purpose, HENS offers many promising improvements over smaller ensembles, including a reduced likelihood of missed events (Figure 11 and 12), a better ability to have ensemble members represent the true value (Figure 6), and better information gain to sample low-likelihood events that occur at the tail of the ensemble distribution (Figure 1,3). These metrics ensure that a large ensemble provides more information than a small one, and they verify that the ensemble does not collapse where each member provides duplicative information. The trajectories that correctly capture the true outcome can be used to study the dynamics and drivers of extremes: a similar process has been used in Mo et al. [2022], Millin and Furtado [2022], Leach et al. [2024]. It can be used to study counterfactuals, such as the probability of avoiding the 2023 Kansas City heatwave (Figure 5 HENS Part II). And it can be used to create a large dataset with many samples of events that exist at the tail of the IFS distribution, which would only have limited samples of the event or would be missed entirely.

But in this case, we would need to still show that the ensemble is actually representing the tails of the distribution correctly. This should be evaluated but it is a very hard problem, not only for the ensemble system, but also for the evaluation as you would need a very long test period to sample extreme events to understand the real quality of the ensemble when representing a 4-sigma event for, say, precipitation with enough statistics. This may not be possible without overlap between training and testing datasets.

We fully agree that there is limited observational data. This is a fundamental constraint. In the face of these limits, we note that we validate HENS against IFS on extreme diagnostics to the maximal extent possible on the time periods available. We hope that HENS, as well as future model improvements, lead to a model with comparable trustworthiness as physics-based models.

We validate HENS on the tails of the observational distribution. This validation is limited by the length of the observational dataset. But we also validate on HENS on its ability to sample the tails of the forecast distribution. For instance, HENS gives us a better estimate of the 99th percentile of all the ensemble members. This test of HENS does not require a large observational dataset, but rather requires comparing the large ensemble to the small ensemble.

HENS represents a first-of-its-kind experiment. If it yields promising results in understanding extreme statistics and drivers, perhaps it can be used as motivation for the weather and climate community to invest in huge ensembles of physics-based simulations. At the very least, if HENS were trained on purely simulated data, this would enable saving all the observed data for validation, and there could also be other perfect model experiments, in which HENS is validated against a large set of physics-based simulations.

If you represent all possible weather situations at day 10, this can well indicate that your model is all over the place when it is basically uncorrelated with the real-world trajectory.

Thank you for raising this point: this is a crucial aspect to validate. The HENS CRPS is similar to the 58-member SFNO-BVMC CRPS at day 4, indicating that the huge ensemble is providing skillful forecasts at early lead times (Figure 8a). The same is also true for the HENS spread-error ratio (Figure C2). If the HENS forecasts were entirely unreliable, then we would expect these scores to be significantly degraded. We will also provide HENS ensemble mean RMSE scores and include calculation of statistics at earlier lead times, as discussed below.

It would be a much stronger statement if you see the same at day 2 or 5. It also smells a bit like cherry-picking when the evaluation is focusing on day 10+ as you see a good spread-error ratio here. I would like to see evaluations of earlier lead times (in particular for Figure 4, 6, 7).

Thank you very much for this feedback. We greatly appreciate this comment, as it will help strengthen our paper. We will perform this analysis at an earlier lead time. We note that while the spread-error ratio is at 1 at day 10, it's reasonably close to 1 at earlier lead times (Figure C2). In our revised manuscript, will also include the HENS ensemble mean RMSE and spread as a function of lead time. For these two reasons, the results from Figure 10 and Figure 11 hold up at earlier lead times (see Figure D1 and Figure 11 itself.)

We emphasize that we are not cherry-picking a lead time of 10 days. We choose to validate this lead time because of our scientific interest in counterfactual trajectories. Forecasts at 10 day lead times are still correlated with the initial conditions, and they are still, on average, more skillful than climatology. At this lead time, we can look at a large trajectory of possible future weather states, accounting for synoptic-scale uncertainty. Still, at 10-days, we can rigorously validate that the results are realistic using medium-range weather diagnostics. For lead times longer than a few weeks, the chaotic limit of predictability requires other climatological diagnostics to be used, since the model is no longer conditioned on the initial conditions. Since ML emulators are new, we chose to focus on medium-range diagnostics for direct, rigorous validation that the ensemble output is trustworthy.

It would also be good if you could show results for more challenging quantities such as precipitation as well.

Thank you for raising this issue. Precipitation is excluded as a variable because of the challenges in obtaining a global training dataset with high-spatiotemporal resolution. Some ML model groups have a "lack of confidence in the quality of ERA5 precipitation data" [Price et al., 2024] and exclude the precipitation results from the main evaluation [Lam et al., 2023]. In addition to the training dataset challenge, the spatial statistics and long tails that are present in precipitation datasets, in some cases, indicate that further architectural changes are necessary for ML models [Pathak et al., 2022]. Precipitation is not included in the original SFNO [Bonev et al., 2023]. We note that the exclusion of precipitation is a common feature across many data-driven weather prediction models [Bi et al., 2023, Keisler, 2022, Chen et al., 2023a,b, Ramavajjala, 2024, Cachay et al., 2024, Bodnar et al., 2024], many of which are leading models listed on WeatherBench.

The addition of precipitation is very much an important challenge at the forefront of data-driven weather prediction. In future research, we certainly wish to emulate precipitation to forecast LLHI precipitation events and will include it as a variable in our ensemble: however, for this work, we focus on the development of ensembles and study surface temperature events (with other variables forthcoming) to our analysis.

The first part that outlines how a huge ensemble can be run and what hardware is needed is very interesting. However, it would be good if you could put the results a bit better into perspective. The data pipeline that you describe seems to bring a machine of the size of Perlmutter to its limits. A 25 GB/s connection is rather expensive to maintain. This does not go down well with the claim the ML models are orders of magnitude cheaper when compared to conventional models?

Thank you for raising this interesting point of comparison and discussion. Regarding computational cost, we are not considering data transfer. Because ML models generate an ensemble member so quickly (one hour for IFS on 96 CPUs, one minute for SFNO on 1 GPU), it is more feasible to run them in huge ensemble configurations. The data transfer stresses are relevant largely because ML makes it reasonable to create 256 ensemble members simultaneously. For instance, in the data pipeline we describe, it was possible for us to access 256 GPUs to generate 256 ensemble members per minute simultaneously (Section 2.1 of HENS Part II). It would be more challenging to request 24,576 CPUs (96 CPUs per ensemble member * 256 members) at once to create 256 members at one time. Even then, each member would take one hour, not one minute, so the data transfer requirements would not be as high. We agree that these new capabilities require new, fast, and expensive data transfer connections. But now these costs also open new science questions around huge ensemble datasets that were impractical to explore with traditional models.

Would it be possible to compare the huge ensemble also against other ML ensemble systems that are published in the literature?

In future work, we agree it would be interesting to compare huge ensembles from multiple ML architectures. However, the core of this study is to assess the effect of ensemble size. The central analysis necessary to perform this goal is to benchmark our huge ensemble against smaller ensembles (58 members) from the same model, and against IFS. Assessing the performance against other ML ensembles is out of the scope of this study, especially since we do not have access to another huge ML ensemble.

Minor comments: P6: How large is the model if you want to send it around instead of the data?

The SFNO checkpoint model weights are 8.4 GB per checkpoint.

How does climate change enter the discussion around huge ensembles?

We choose to run our huge ensemble in summer 2023, the hottest summer on record [Esper et al., 2024]. This allows us to study alternate trajectories that could have occurred due to internal variability in one of the warmest summers on record. However, we do not explicitly model the influence of CO_2 in our forecasts, as other emulators have recently added this capability [Watt-Meyer et al., 2024].

Figure 11 seems to have an error in the caption with 240, 246, 252… not fitting to day 4,7,10.

We have fixed this error, thank you very much.

Comments from Reviewer #2

In Part 1, the integrated system was validated, while in Part 2, the focus shifted to simulating extreme weather events, particularly those exceeding 4 standard deviations from the mean. The creation and analysis of 7,424 ensemble members using a range of probabilistic metrics is impressive and represents a significant advancement in ensemble-based forecasting. This large ensemble approach has substantial potential for improving the prediction and assessment of extreme weather events, offering valuable insights into their likelihood and associated uncertainties. Moreover, the integration of artificial intelligence, specifically through Spherical Fourier Neural Operators, presents a promising new direction for weather forecasting, combining computational efficiency with robust performance. However, several concerns need to be addressed, as outlined in the detailed comments below. With these revisions, we believe the manuscript will be well-prepared for publication.

Thank you for your review of our paper.

Major Comments 1. The authors do not demonstrate whether the error accumulates and spreads as the lead time progresses, and the explanation for this omission is unclear. Since the perturbations are based on bred vectors, the lack of error accumulation could potentially result from deviations introduced by the initial bred vector itself. Therefore, including an analysis of the variance and characteristics of the bred vector would enhance the validity of the results and provide a more convincing argument.

Thank you for this comment: we will clarify this in our upcoming manuscript. In Figure C2 of our manuscript, we show that the spread-error ratio is approximately 1 for HENS (the 7,424 huge ensemble) and for a 58-member ensemble. This is an important benchmark metric that indicates that the ensemble spread and the ensemble mean RMSE are comparable. We show the ensemble mean RMSE in part I of our manuscript (Figure 9) for the 58-member ensemble; it grows as a function of lead time. In our revised manuscript, we will show that the ensemble mean RMSE and variance (with the reasonable spread-error ratio in Figure C2) for HENS also grows with lead time.

2. The authors have only analyzed temperature, but with the availability of u10m data, further analysis of wind gusts could be conducted. Limiting the results to temperature alone restricts the reliability of the study. Additional analyses of other extreme events using different variables would strengthen the manuscript. If it is feasible to modify the deep learning model to simulate precipitation, I would recommend including it. If not, at the very least, wind gust analysis should be explored and discussed.

Thank you for raising this suggestion. We will include an analysis of wind in our results in our revised manuscript. It is not possible to modify the deep learning model to simulate precipitation (see the manuscript comments for Part I), as that is not one of the variables in the backbone of SFNO and many other data-driven weather prediction models.

3.The study focuses exclusively on heat waves from June to August. However, cold waves represent the opposite end of temperature extremes and are equally important. If the authors can demonstrate that their system is capable of reproducing cold waves, it would significantly enhance the reliability of the model in predicting a broader range of temperature extremes.

Thank you for raising this possibility. We evaluated the cold tail of the ensemble distribution in the calculation of ensemble gain (Eqn F2, Figure 1-3), the large sample behavior of the 0.1st and 10th percentiles

(Figure 4), and in the calculation of the outlier statistic (Figure 11). We believe that this analysis of the cold tails is sufficient for our core scientific questions, and we respectfully suggest that further evaluation of cold waves would not further the core of this manuscript on huge ensembles. We do not intend to further evaluate cold waves in our huge ensemble simulation, as we only have one season of data (June-July-August 2023). This would not be enough data to evaluate cold waves in midlatitude land. We intentionally selected this season to study for very specific reasons because it was one of the warmest seasons on record [Esper et al., 2024], and we believe that further analysis of cold waves could happen in future work, particularly simulations of other seasons.

Minor Comments 1. Perturbations were applied using both bred vectors and checkpoints. One question that arises is which of these two methods is more sensitive to an increased perturbation. A brief analysis and discussion on this point would be valuable for readers to better understand the relative impact of each approach.

We compare the relative influence of multi-checkpointing and bred vectors in our Part I manuscript, Figure 6. We discuss this behavior in Part I as it is a fundamental question about our ensemble setup, rather than something intrinsic in the HENS run in summer 2023, which is the main topic of Part II. Our current discussion on the topic is as follows:

As a model perturbation, multi-checkpointing does not represent the uncertainty arising from an imperfect initial condition. Therefore, the multi-checkpoint ensemble is underdispersive at early lead times. On the other hand, the ensemble composed only of bred vectors is underdispersive on synoptic time scales (3-5 days) when representing model uncertainty also becomes important for obtaining good calibration.

Furthermore, we show that multi-checkpointing (beyond the 29 checkpoints we use) is not sensitive to increased perturbation. Increasing the multi-checkpointing perturbation would mean creating an ensemble with more than 29 checkpoints. As a function of the number of checkpoints, we show that the ensemble spread has converged (for t2m in the current manuscript, and with other variables in the next version) by 29 checkpoints (Figure 3, HENS Part I).

2. There are typos in plural and singular, please find and correct them.

Thank you. We will do so.

References

K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi:10.1038/s41586-023-06185-3. URL http://dx.doi.org/10.1038/s41586-023-06185-3. 4

C. Bodnar, W. P. Bruinsma, A. Lucic, M. Stanley, A. Vaughan, J. Brandstetter, P. Garvan, M. Riechert, J. A. Weyn, H. Dong, J. K. Gupta, K. Thambiratnam, A. T. Archibald, C.-C. Wu, E. Heider, M. Welling, R. E. Turner, and P. Perdikaris. A foundation model for the earth system, 2024. URL https://arxiv.org/abs/2405.13063. 4

B. Bonev, T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar. Spherical Fourier Neural Operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023. 4

S. R. Cachay, B. Henn, O. Watt-Meyer, C. S. Bretherton, and R. Yu. Probabilistic Emulation of a Global Climate Model with Spherical DYffusion. *arXiv preprint arXiv:2406.14798*, 2024. 4

K. Chen, T. Han, J. Gong, L. Bai, F. Ling, J.-J. Luo, X. Chen, L. Ma, T. Zhang, R. Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a. 4

L. Chen, X. Zhong, F. Zhang, Y. Cheng, Y. Xu, Y. Qi, and H. Li. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023b. doi:10.1038/s41612-023-00512-1. URL https://doi.org/10.1038/s41612-023-00512-1. 4

J. Esper, M. Torbenson, and U. Büntgen. 2023 summer warmth unparalleled over the past 2, 000 years. *Nature*, 631(8019):94–97, May 2024. ISSN 1476-4687. doi:10.1038/s41586-024-07512-y. URL http://dx.doi.org/10.1038/s41586-024-07512-y. 5, 7

R. Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022. 4

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382 (6677):1416–1421, 2023. doi:10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336. 4

N. J. Leach, C. D. Roberts, M. Aengenheyster, D. Heathcote, D. M. Mitchell, V. Thompson, T. Palmer, A. Weisheimer, and M. R. Allen. Heatwave attribution based on reliable operational weather forecasts. *Nature Communications*, 15(1), May 2024. ISSN 2041-1723. doi:10.1038/s41467-024-48280-7. URL http://dx.doi.org/10.1038/s41467-024-48280-7. 3

A. Mahesh, W. Collins, B. Bonev, N. Brenowitz, Y. Cohen, J. Elms, P. Harrington, K. Kashinath, T. Kurth, J. North, et al. Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators. *arXiv preprint arXiv:2408.03100*, 2024a. 1

A. Mahesh, W. Collins, B. Bonev, N. Brenowitz, Y. Cohen, P. Harrington, K. Kashinath, T. Kurth, J. North, T. O'Brien, et al. Huge Ensembles Part II: Properties of a Huge Ensemble of Hindcasts Generated with Spherical Fourier Neural Operators. *arXiv preprint arXiv:2408.01581*, 2024b. 1

O. T. Millin and J. C. Furtado. The Role of Wave Breaking in the Development and Subseasonal Forecasts of the February 2021 Great Plains Cold Air Outbreak. *Geophysical Research Letters*, 49(21), Oct. 2022. ISSN 1944-8007. doi:10.1029/2022gl100835. URL http://dx.doi.org/10.1029/2022GL100835. 3

R. Mo, H. Lin, and F. Vitart. An anomalous warm-season trans-Pacific atmospheric river linked to the 2021 western North America heatwave. *Communications Earth and Environment*, 3(1), June 2022. ISSN 2662-4435. doi:10.1038/s43247-022-00459-w. URL http://dx.doi.org/10.1038/s43247-022-00459-w. 3

J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 4

I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, Dec. 2024. ISSN 1476-4687. doi:10.1038/s41586-024-08252-9. URL http://dx.doi.org/10.1038/s41586-024-08252-9. 4

V. Ramavajjala. HEAL-ViT: Vision Transformers on a spherical mesh for medium-range weather forecasting. *arXiv preprint arXiv:2403.17016*, 2024. 4

O. Watt-Meyer, B. Henn, J. McGibbon, S. K. Clark, A. Kwa, W. A. Perkins, E. Wu, L. Harris, and C. S. Bretherton. Ace2: Accurately learning subseasonal to decadal atmospheric variability and forced responses, 2024. URL https://arxiv.org/abs/2411.11268. 5