AURORA 1.5: FINE-TUNING A FOUNDATION MODEL FOR MEDIUM-RANGE ENSEMBLE WEATHER PREDICTION

A PREPRINT

Jonathan A. Weyn[1,*], Zekun Ni[1], Amit Misra[2], Will Fein[2], Haiyu Dong[1], Wessel P. Bruinsma[4], Richard E. Turner[4], Matt Corey[1], Kit Thambiratnam[1], Kevin White[2], Kenji Takeda[3], Hongyu Sun[1]

[1] *Microsoft Corporation, Redmond, Washington, USA*

[2] *Microsoft AI for Good Lab, Redmond, Washington, USA*

[3] *Microsoft Research Accelerator, Cambridge, UK*

[4] *University of Cambridge, Cambridge, UK*

July 9, 2026

ABSTRACT

We present **Aurora 1.5**, a fine-tuned variant of the Aurora atmospheric foundation model [Bodnar et al., 2025] optimized for skillful medium-range ensemble weather prediction. Building on Aurora's pretraining across diverse heterogeneous atmospheric data, we introduce a three-stage fine-tuning pipeline that (i) expands the single-level variable set and randomizes lead-time embeddings to enable a native one-hour temporal resolution, (ii) injects Gaussian noise into AdaptiveLayerNorm (AdaLN) modules to generate stochastic forward passes and optimizes a Continuous Ranked Probability Score (CRPS) objective in place of a deterministic loss, and (iii) performs auto-regressive fine-tuning on operational ECMWF analyses with multi-step rollouts. Aurora 1.5 ENS outperforms the ECMWF ENS operational ensemble on 88.9% of upper-air and single-level target variables in the medium range (days 1–10). Reliability diagnostics including rank histograms indicate that Aurora 1.5 ENS achieves this result with a slight over-dispersion, in contrast to the under-dispersion typical of dynamical models. Compared to Aurora, Aurora 1.5 also better predicts extreme events, reducing tropical cyclone track errors by 16% and the mean absolute error on top-5$^{\text{th}}$ percentile heat waves by 58%. Aurora 1.5 demonstrates that foundation model fine-tuning is a viable, cost-effective path toward reliable probabilistic AI weather prediction.

Keywords ensemble weather prediction · foundation models · uncertainty quantification · deep learning · AI weather forecasting

1 Introduction

Operational medium-range weather forecasting relies on ensembles of numerical weather prediction (NWP) simulations to characterize forecast uncertainty arising from initial-condition errors and model errors [Leutbecher and Palmer, 2008, Palmer, 2019]. These probabilistic forecasts are crucial for many downstream applications such as extreme weather mitigation, energy and resource management, and logistics.

Recent advances in data-driven weather prediction have produced deterministic models that exceed the skill of leading operational NWP systems for many variables and lead times [e.g., Lam et al., 2023, Bi et al., 2023, Lang et al., 2024]. In particular, Aurora, a foundation model for the atmosphere, has demonstrated the ability of a generalized model pre-trained on a large, diverse set of data to exhibit strong transfer behavior to new prediction tasks even under a relatively light fine-tuning regimen [Bodnar et al., 2025]. Meanwhile, probabilistic methods have been used to carry these benefits to stochastic weather predictions, using generative approaches as in GenCast [Price et al., 2024] or noise-injection approaches such as that used in the AIFS-CRPS [Lang et al., 2026] to produce ensembles of forecasts.

In this paper, we present **Aurora 1.5**, a fine-tuned variant of the Aurora foundation model [Bodnar et al., 2025] that brings several notable improvements for medium-range weather prediction, including a stochastic ensemble variant, Aurora 1.5 ENS. Specifically, our contributions are:

- Expansion of the single-level variable set, enabling Aurora to predict useful parameters such as cloud cover, precipitation, and radiation;
- Utilization of randomized lead-time embeddings at training time to enable a native 1-hour temporal resolution and improve long-range stability;
- A fine-tuning stage enabling stochastic ensemble forecasts with model perturbations.

Our fine-tuning recipe produces two model variants, a deterministic Aurora 1.5 and a probabilistic Aurora 1.5 ENS. For the probabilistic model, given the modest changes required, we choose the noise-injection method inspired by the AIFS-CRPS: inserting stochasticity into the network via noise injection into AdaptiveLayerNorm (AdaLN) modules [Perez et al., 2018, Peebles and Xie, 2023], running multiple forward passes, and directly minimizing the Continuous Ranked Probability Score [CRPS, Gneiting and Raftery, 2007]. Aurora 1.5 consistently outperforms the state-of-the-art best variant of the original Aurora model [Bodnar et al., 2025] while also outperforming the European Centre for Medium-range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) dynamical model on newly-added variables such as total cloud cover. Meanwhile, Aurora 1.5 ENS outperforms ECMWF's dynamical ensemble ENS on 88.9% of upper-air and single-level targets. We also demonstrate the applicability of Aurora 1.5 to extreme weather forecasting, focusing on tropical cyclone tracks and extreme heat and cold events.

2 The Aurora 1.5 model

Aurora is a 3D Swin-Transformer-based foundation model of the atmosphere, pretrained on a large and diverse corpus of atmospheric data and with a demonstrated ability to excel at downstream prediction tasks such as ocean waves and atmospheric chemistry [Bodnar et al., 2025]. The model follows an encoder–processor–decoder design in which a flexible Perceiver-style encoder maps an arbitrary set of input variables onto a shared latent representation, a transformer backbone advances this representation in time, and a decoder reconstructs predicted variables on the native latitude–longitude grid. Some design features of Aurora are particularly useful for the extensions herein:

1. **Variable flexibility.** The encoder and decoder learn per-variable embeddings, allowing new variables to be added with minimal architectural change and with most pretrained weights reused. The Perceiver architecture combines variable embeddings additively, making it easy to add in new variable parameters.
2. **Lead-time conditioning.** The backbone contains a Fourier feature embedding of the prediction lead time, unused in the deterministic pretrained model. We leverage this mechanism for training at arbitrary, randomized lead times. See Section 6.3 for more details on model rollouts with the variable lead time.
3. **Layer normalization modules.** The existing presence of LayerNorm modules in the backbone make it easy to inject the stochastic noise.

Aurora 1.5 is trained across a few stages which are summarized here. More details can be found in Section 6.

1. **Stage 1: More variables and multiple lead times.** Using hourly ERA5 as inputs and targets, we train starting from the Aurora pretrained checkpoint with a reduced learning rate of 5×10^{-5} and cosine annealing. New parameters used in the lead time embedding and the encoder/decoder layers for the new variables have a higher learning rate. We train on ERA5 data from 1981–2023 and validate with 1979–1980, although in practice, validation loss always continued to decline with the training loss. Mean absolute error (MAE) is kept as the optimization target and reasonable scaling parameters are added for the new variables. The gap between the last input time step and the target time step is randomized in 1-h increments from 0 h to 12 h with higher likelihood of selecting 6 h and 12 h.
   (a) **Stage 1a: Auto-regressive fine-tuning.** The model is next trained on a two-step auto-regressive strategy as was the original Aurora model. Gradients are computed through the entire two steps. The learning rate is reduced to 3×10^{-6} and held constant. Because of the auto-regressive requirement, multiple lead times are not used.
   (b) **Stage 1b: Analysis fine-tuning.** To better fit Aurora to real-time data from IFS initial conditions, we do a short auto-regressive fine-tune as in the previous step but with IFS analysis data from 2018-2023. This stage produces the model we refer to as Aurora 1.5, the deterministic version.

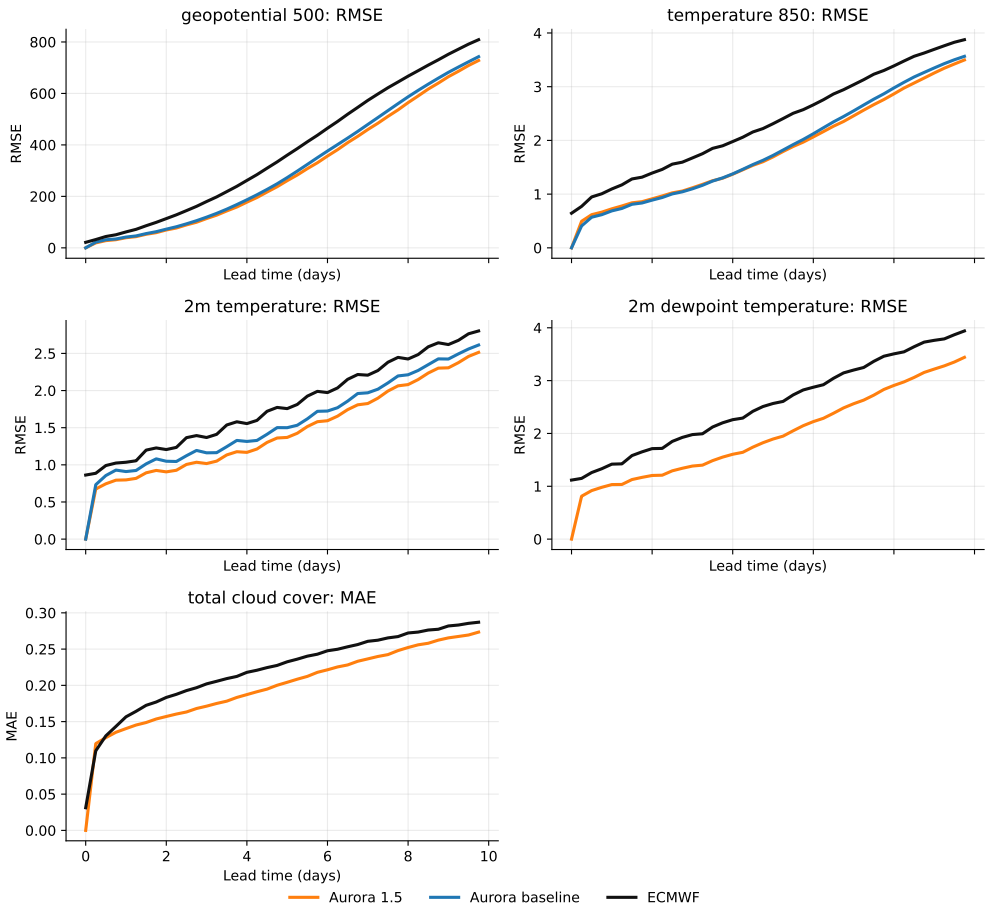

Figure 1: **Deterministic verification of Aurora 1.5 against the IFS HRES and Aurora baseline forecasts**, measured against IFS analysis. Lines indicate RMSE or MAE in native variable units as a function of forecast lead time. Across highlighted variables, including some new in Aurora 1.5, Aurora 1.5 significantly improves on the dynamical ECMWF IFS model. Aurora 1.5 even outperforms the Aurora baseline model on variables predicted by both variants.

2. **Stage 2: Stochastic fine-tuning.** The Stage 1 model is next fine-tuned while adding the noise embeddings to the backbone's AdaLN. We use cosine annealing with a maximum learning rate of 5×10^{-5} while the noise modulation parameters have a rate of 2×10^{-4}. The optimization is switched to the simplified 2-member CRPS formulation as each iteration step runs two forward passes with which to compute the loss.
   (a) **Stage 2a: Auto-regressive stochastic fine-tuning.** As in Stage 1a, we then apply auto-regressive fine-tuning to improve rollout accuracy. Unlike 1a, we rollout over 4 steps covering a 24-hour forecast. Due to memory limitations a stop-gradient method is used which does not propagate gradients through the entire rollout.
   (b) **Stage 2b: Analysis fine-tuning.** As in Stage 1b, we fine-tune with the same configuration as 2a but with analysis data. This stage produces the model we refer to as Aurora 1.5 ENS.

3 Results

3.1 Improved deterministic forecasts from Aurora 1.5

Stage 1 fine-tuning yields consistent improvements over the Aurora baseline[1] across single-level variables, while upper-air variables retain the strong skill of the pretrained model. This result is shown in Fig. 1, which averages

[1]Note that this evaluation uses an Aurora baseline specialized with a 2-step auto-regressive fine tune on top of the Aurora 0.25° Fine-tuned checkpoint that is the recommended public model version. We disable LoRA throughout.

root-mean-squared error (RMSE) and MAE computed over forecasts initialized every Monday and Thursday 00 UTC during the test year 2024 (103 total forecast issue times). For new variables such as cloud cover, Aurora 1.5 continues to significantly outperform the ECMWF IFS baseline. We also note that, in ablation experiments, the randomized lead-time training improves long-rollout stability (not shown).

3.2 Ensemble verification against ECMWF ENS

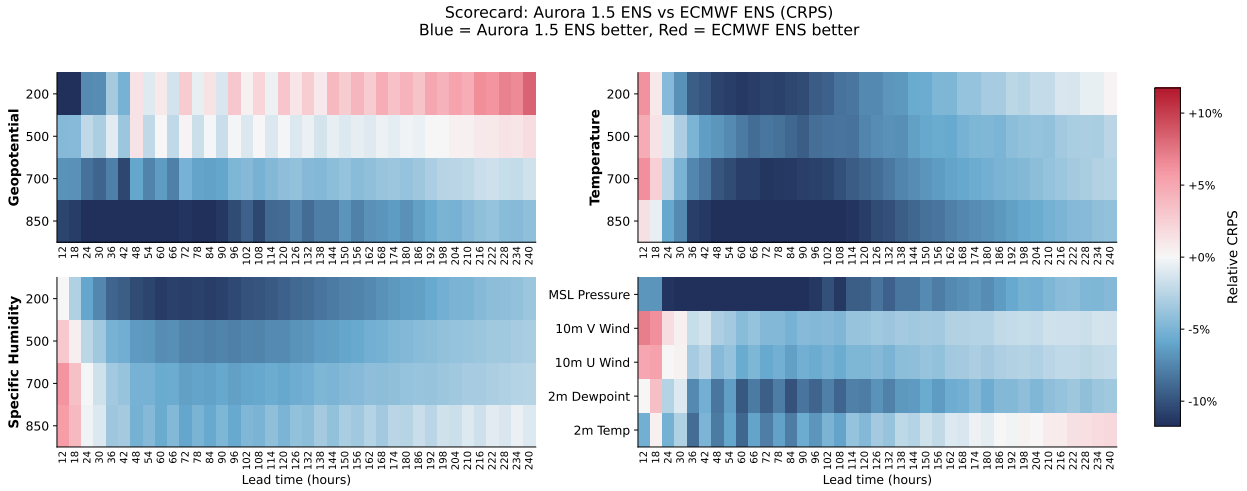

Figure 2: **Probabilistic verification of Aurora 1.5 ENS against ECMWF ENS.** Shading indicates relative CRPS values with ECMWF ENS as the baseline. Blue (red) colors show where Aurora 1.5 ENS performs better (worse) than the baseline. Across upper-air geopotential, temperature, and humidity, and five near-surface variables, Aurora 1.5 ENS outperforms ECMWF ENS on 88.9% of variable-step targets.

Across a range of upper-air and near-surface variables and up to 10-day forecasts, Aurora 1.5 ENS improves upon the CRPS score of the baseline ECMWF ENS on 88.9% of variable- and lead-time- targets. These results are shown in Fig. 2, where we use the relative difference in CRPS between Aurora 1.5 ENS and the ECMWF ENS. The test set used to evaluate the ensemble forecasts contains the same 103 forecast issue times, but we note that Aurora 1.5 ENS was initialized using the initial-condition perturbations present in ECMWF ENS by using steps 0 h and 6 h as the two input steps to the model. This choice was made for both practical (readily available data) and tactical (ensuring physically-consistent evolution of the initial-condition perturbations between the two input steps) reasons, but it means that Aurora 1.5 ENS may not quite generalize as well to this flavor of input data. We also show the scorecards starting at only step 12 h in Fig. 2 for this reason. All 50 members are used for both models.

The initialization may explain why temperature, specific humidity, and a few near-surface variables are initially marginally worse than the baseline over the first 18 hours, but subsequently, Aurora 1.5 ENS comfortably outperforms ECMWF ENS with only a few exceptions: 2-m temperature is slightly worse by day 10, while geopotential at 200 hPa is generally worse. Reliability of the ensemble is assessed using rank histograms [Hamill, 2001], shown in Fig. 7. Typically, Aurora 1.5 ENS achieves its improved CRPS targets by greatly increasing the ensemble spread compared to ECMWF ENS; in practice, this is often desirable since dynamical models typically suffer from under-dispersion, although in our case, it is over-compensated for as the ensemble becomes over-dispersive. We also see some possible explanations for the 2-m temperature and 200-hPa geopotential: both tend to have a low bias, while temperature also has less spread compared to other variables.

Another benefit of the noise-injection model with CRPS loss is that the model is free to retain physical structures without the smoothing characteristic of deterministic models trained with only regression loss. An example output illustrating the capability of Aurora 1.5 ENS is shown in Fig. 3, where we plot ensemble mean and standard deviation of cloud cover and downwards solar radiation at the surface. The model reasonably shows greater uncertainty in areas with likely variations in cloud cover due to localization of features such as low pressure waves. Aurora 1.5 ENS is able to retain much more detailed physical fields compared to Aurora 1.5 (not shown).

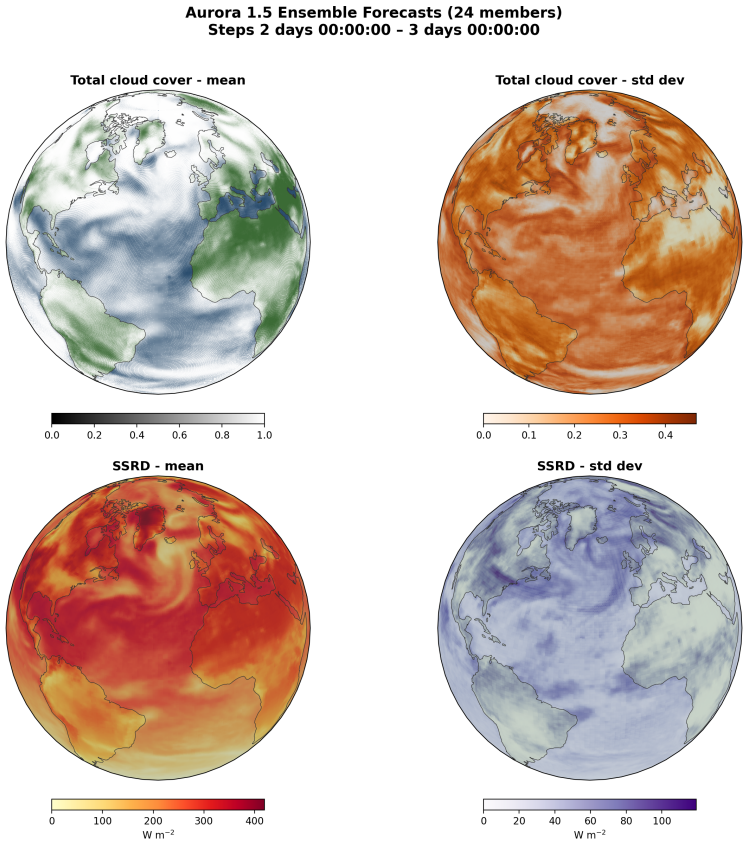

Figure 3: **Example Aurora 1.5 ENS forecast.** The outputs are from a single Aurora 1.5 ENS forecast issued at 18Z 24 June 2026 and averaged over lead times 48–72 h, showing two new single-level variables (total cloud cover, downward solar radiation at the surface in one hour) and their characteristic fields. Mean and standard deviation are shown as an example of possible outputs from the ensemble.

4 Forecasting extreme weather

The improvements made in Aurora 1.5 and the availability of the ensemble prediction system of Aurora 1.5 ENS translate to notable improvements in forecasting extreme events using a global data-driven weather model. We document applications to tropical cyclone track forecasting and extreme heat and cold events herein. To capture more initialization times for these events, the test set differs for this section: forecasts are issued four times daily; forecasts are issued throughout the years 2024–2025; Aurora 1.5 ENS is reduced to 32 members; and all members are initialized with ECMWF IFS control T0 data, without initial condition perturbations. This latter choice is made for practical reasons only given the time that would be needed to obtain the specific set of initial conditions from ECMWF's MARS archival system. Anecdotally, the ensemble spread may be slightly lower at early lead times, but the model perturbations soon dominate.

4.1 Tropical cyclone track forecasts

As shown in Bodnar et al. [2025], Aurora achieved state-of-the-art accuracy on tracking tropical cyclones, outperforming all operational forecasting centers even with a deterministic model. To evaluate Aurora 1.5, we compute tropical cyclone track forecasts over 2024–2025 and measure against IBTrACS [Knapp et al., 2010] best-track positions, comparing them with the original Aurora. Forecasts are initialized four times daily and evaluated on a 6-hour cadence over storm-centered forecast windows. Figure 4 shows that Aurora 1.5 reduces track error relative to Aurora by 9–24% across lead days 1–5. The ensemble further improves on Aurora by 13–34% when aggregating with the median track location, which performs better than aggregating with the mean location.

As an aggregate measure of ensemble reliability, we show track location rank histograms in Figure 8 for Aurora 1.5 ENS and ECMWF ENS. These histograms are computed by ranking the true cyclone position relative to individual members

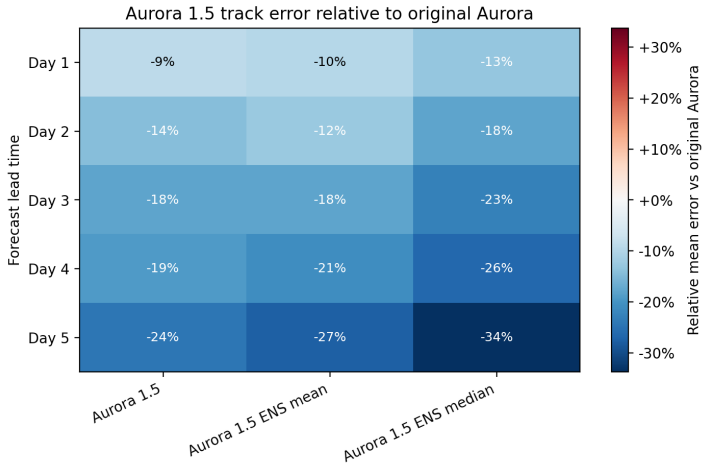

Figure 4: **Aggregate Aurora 1.5 tropical cyclone track error improvement.** Cells show the percent change in mean great-circle track error relative to original Aurora by forecast lead day. Negative values indicate lower error for Aurora 1.5.

and centered around the ensemble centroid, with a normalized rank of 1.0 indicating that the true position lies entirely outside the ensemble cloud. While due to the lack of initial-condition perturbations there is a bias towards rank 1.0 in 1-day forecasts from Aurora 1.5 ENS, by day 3, the tracks are very well-distributed, while those of ECMWF ENS still tend to be over-dispersed.

Figure 5 shows an illustrative forecast for Hurricane Helene. For this forecast, the National Hurricane Center (NHC) official forecast has mean track error 110.6 km, while the Aurora 1.5 forecast has mean error 58.4 km and the Aurora 1.5 ENS mean has mean error 26.3 km. Both the control and ensemble mean accurately predict the landfall along the Florida panhandle and the spread of ensemble members captures the true track position for the duration of the system.

4.2 Extreme temperature events

We also test Aurora 1.5's ability to predict extreme heat waves and cold snaps in the medium range by evaluating 2-meter temperature forecasts over 2024–2025 against ERA5 daily maximum ($t_{\max}$) and minimum ($t_{\min}$) temperatures at lead times between 5-6 days (forecast hours 126, 132, 138, and 144). The event sets and RMSE curves are computed from a random sample across all locations and all 00 UTC initialization times within the test period.

First, a threshold-conditioned RMSE diagnostic asks how forecast error changes as verification samples move into the warm and cold tails of the local temperature distribution. For a sweep of climatological z-score thresholds, RMSE is computed on samples warmer than the threshold for positive z-scores and colder than the threshold for negative z-scores. The plotted quantity in Figure 6 is relative RMSE versus original Aurora[2], so negative values indicate better forecasts (lower error than the baseline). Second, Table 1 reports CRPS for the full ensemble distribution on four extreme-temperature event classes: daily $t_{\max}$ above the annual p90 and p95 thresholds, and daily $t_{\min}$ below the annual p05 and p10 thresholds. The percentile thresholds are computed per grid cell from the 1991–2020 ERA5 climatology and applied to the 2024–2025 verification period.

Figure 6 shows that Aurora 1.5 improves the warm-tail $t_{\max}$ forecasts relative to original Aurora across much of the z-score range. Even the deterministic Aurora 1.5 forecast outperforms the original-Aurora baseline by 10% or more, with improvements of more than 45% on the hot tail. Aurora 1.5 ENS meanwhile uses the ensemble information to further improve on the deterministic score. The $t_{\min}$ result is more nuanced: while the ensemble median outperforms the original Aurora deterministic forecast, the deterministic Aurora 1.5 forecast is less consistent, with a few percentage points worse performance on the abnormally cold events. Overall, the threshold-conditioned RMSE view is the clearest summary of point-forecast changes across the temperature distribution: Aurora 1.5 improves hot-tail $t_{\max}$ most clearly, while cold-tail $t_{\min}$ remains the harder case.

[2]This evaluation is based on the original Aurora 0.25° Pre-trained, as we found it to perform better than the IFS-optimized Aurora 0.25° Fine-tuned.

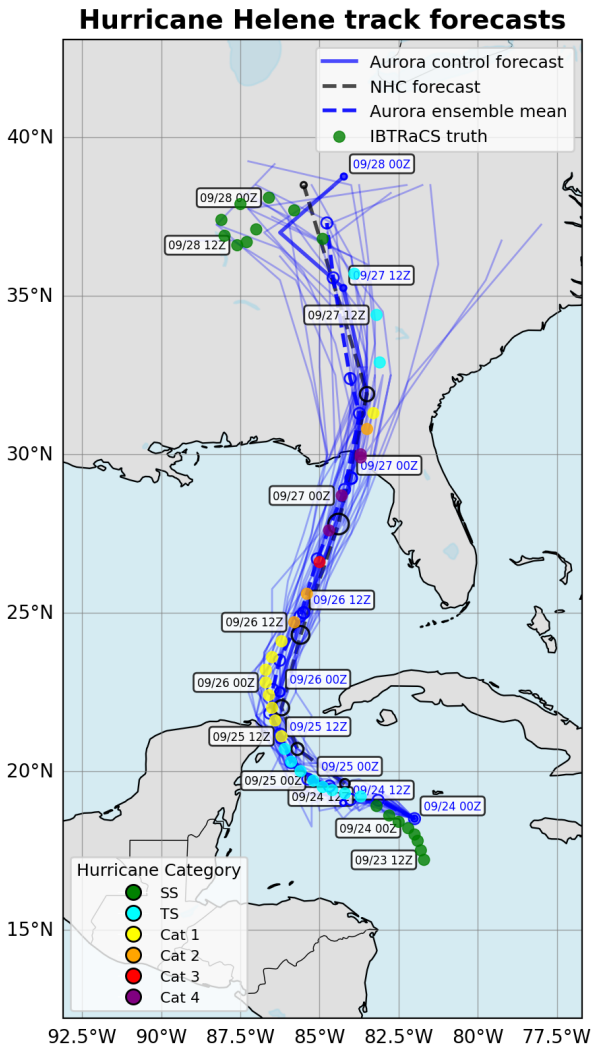

Figure 5: **Aurora 1.5 forecasts for Hurricane Helene.** An example forecast issue at 00 Z 24 September 2024 shows how forecasts from Aurora 1.5 and Aurora 1.5 ENS compare to the verified best track and the official National Hurricane Center (NHC) guidance forecast. A subset (24 of 32) of the individual members of Aurora 1.5 ENS effectively capture every verifying observation after accounting for the identical initial location in each member. While approximately 6 hours too fast, the ensemble mean and control forecast tracks accurately predict the location of Helene's landfall in Florida.

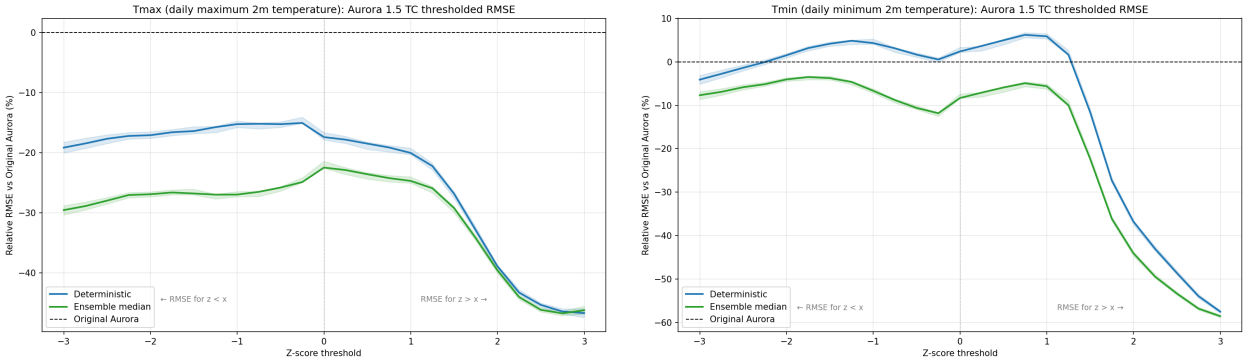

Figure 6: **Thresholded RMSE for daily t_{max} and t_{min} .** The left panel shows daily t_{max} and the right panel shows daily t_{min} . Relative RMSE is shown versus original Aurora. Negative values indicate lower RMSE than original Aurora. The blue line is Aurora 1.5; the green line is the Aurora 1.5 ENS 32-member ensemble median.

Table 1: **CRPS for the Aurora 1.5 ENS 32-member ensemble on extreme-temperature events.** Lower is better. For a deterministic forecast, CRPS reduces to MAE, so original Aurora MAE is included as a deterministic reference.

Event class	Event count	Orig. MAE	A1.5 median MAE	A1.5 CRPS	95% CI
$t_{\text{max}} >$ p90	7,661,227	1.703	1.000	0.733	[0.730, 0.738]
$t_{\text{max}} >$ p95	4,592,688	1.795	1.034	0.761	[0.756, 0.764]
$t_{\text{min}} <$ p10	2,862,642	2.229	2.226	1.675	[1.668, 1.684]
$t_{\text{min}} <$ p05	1,221,630	2.379	2.487	1.890	[1.882, 1.898]

Table 1 reports the probabilistic score for Aurora 1.5 ENS. While CRPS reduces to MAE for a deterministic forecast, meaning that the original Aurora MAE provides a usable reference scale, it should not be considered a fair comparison. In fact, for cold events, the ensemble median performs slightly worse on MAE than the original Aurora. However, when accounting for the full distribution, the CRPS of Aurora 1.5 ENS is well below original Aurora MAE for each event class, indicating that the probabilistic ensemble forecast adds value relative to the deterministic baseline under this proper scoring rule. The hot-event CRPS values are around 0.73–0.76 K, while the cold-event CRPS values are larger, around 1.67–1.89 K. This suggests that the ensemble distribution is more accurate for warm-tail t_{max} events than for cold-tail t_{min} events, matching the RMSE picture that cold extremes are more challenging.

5 Discussion

We have presented Aurora 1.5, a fine-tuned variant of the Aurora atmospheric foundation model that delivers both improved deterministic forecasts and skillful medium-range ensemble predictions. Through a multi-stage fine-tuning pipeline – expanding the variable set and temporal resolution, injecting structured noise into AdaLN modules with a CRPS training objective, and auto-regressive fine-tuning on operational analyses – Aurora 1.5 ENS outperforms the ECMWF ENS operational ensemble on 88.9% of upper-air and single-level targets over days 1–10. We further demonstrated that these improvements extend to extreme weather forecasting, with tropical cyclone track errors reduced by up to 24% relative to Aurora (34% for the ensemble median) and substantially lower errors for warm-tail temperature extremes. Our key conclusions are as follows.

Foundation-model fine-tuning is a viable path to probabilistic prediction. A modest fine-tuning recipe, applied to a pretrained foundation model, can produce a competitive ensemble system at a fraction of the cost of training a probabilistic model from scratch. The pretrained Aurora backbone provides a strong initialization that already captures the dynamics of the atmosphere across a wide range of scales; fine-tuning then needs only to introduce stochasticity, optimize a probabilistic objective, and adapt the model to operational initial conditions. This supports the broader hypothesis that the value of atmospheric foundation models lies not only in deterministic skill but in the leverage they provide for downstream tasks such as ensemble forecasting, extended-range prediction, and domain-specific applications.

Probabilistic AI weather models improve extreme event forecasting. The application to tropical cyclone tracks and extreme temperature events demonstrates that the improvements in Aurora 1.5 are not limited to aggregate medium-

range scores. For tropical cyclones, the ensemble median track provides a 13–34% reduction in track error over Aurora, which already outperformed all operational guidance forecasts, and rank histograms indicate well-calibrated spread by day 3. For extreme temperatures, Aurora 1.5 most clearly improves warm-tail $t_{\max}$ forecasts, with ensemble CRPS values 57% below the deterministic Aurora MAE for events exceeding the 95th climatological percentile. Cold-tail $t_{\min}$ remains a harder problem, consistent with a residual warm bias in the model. Bias calibration is a potential avenue for more quick gains on temperature metrics.

Limitations and future work. A few minor issues and improvements are worth pointing out. Aurora 1.5 ENS struggles compared to the ECMWF NWP model on geopotential, especially higher in the atmosphere, possibly because geopotential is already very well-predicted by the dynamical models with little room for improvement. Our model tends to have a low bias, as seen in the rank histograms in Fig. 7. At early lead times, the CRPS of Aurora 1.5 ENS tends to be worse, although this may be an artifact of the data chosen for initialization. However, we do acknowledge that the AI model learns to optimize CRPS by increasing the spread to much higher spread-skill ratios compared to ECMWF-ENS, especially at early lead times. Finally, this study does not consider alternative methods of adding stochastic model perturbations including diffusion models [Price et al., 2023]. A comprehensive cross-model validation should be conducted but is currently out of scope.

Acknowledgments

We thank the original Aurora research team for developing the underlying model, datasets, infrastructure, and scientific insights on which this work builds. We also thank the European Centre for Medium-Range Weather Forecasts (ECMWF) for their commitment to open science and their substantial efforts to generate, curate, and share the datasets that enabled our work, and Matthew Chantry for assistance with data. Finally, we also thank more team members at Microsoft for contributing feedback, including Shuang Qin, Divya Kumar, Gareth O'Brien, and Kai Neuffer.

Data and code availability

ERA5 reanalysis is publicly available from the Copernicus Climate Change Service. ECMWF operational HRES analyses and ENS forecasts are available to authorized users under ECMWF's data policies. Aurora model code and checkpoints, including Aurora 1.5 and Aurora 1.5 ENS, are available at `https://github.com/microsoft/aurora` and links therein.

Competing interests

The authors declare no competing interests.

References

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, May 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL `https://www.nature.com/articles/s41586-025-09005-y`.

M. Leutbecher and T. N. Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227(7):3515–3539, March 2008. ISSN 0021-9991. doi: 10.1016/j.jcp.2007.02.014. URL `https://www.sciencedirect.com/science/article/pii/S0021999107000812`.

Tim Palmer. The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145(S1):12–24, 2019. ISSN 1477-870X. doi: 10.1002/qj.3383. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3383`. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3383.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, December 2023. doi: 10.1126/science.adi2336. URL `https://www.science.org/doi/abs/10.1126/science.adi2336`.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06185-3. URL https://www.nature.com/articles/s41586-023-06185-3.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system, August 2024. URL http://arxiv.org/abs/2406.01465. arXiv:2406.01465 [physics.ao-ph].

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, pages 1–7, December 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9. URL https://www.nature.com/articles/s41586-024-08252-9.

Simon Lang, Mihai Alexe, Mariana CA Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D Dueben, Sara Hahner, et al. Aifs-crps: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1):18, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. pages 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437. _eprint: https://doi.org/10.1198/016214506000001437.

Thomas M. Hamill. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, 129(3):550–560, March 2001. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

Kenneth R Knapp, Michael C Kruk, David H Levinson, Howard J Diamond, and Charles J Neumann. The international best track archive for climate stewardship (ibtracs) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Timo Ewalds, Andrew El-Kadi, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. GenCast: Diffusion-based ensemble forecasting for medium-range weather, December 2023. URL http://arxiv.org/abs/2312.15796. arXiv:2312.15796 [physics].

Fanny Lehmann, Firat Ozdemir, Yun Cheng, Torsten Hoefler, Sebastian Schemm, Benedikt Soja, and Siddhartha Mishra. Can ai weather models predict beyond two weeks? a quantitative benchmark and analysis of long rollouts. *arXiv preprint arXiv:2605.30184*, 2026.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

6 Methods

6.1 More training details

Table 2 shows more details on the training configurations for Aurora 1.5 and Aurora 1.5 ENS.

Table 2: **Model training stage parameters.** The type of GPUs used are NVIDIA A100 80GB.

	Stage 1	**Stage 1a**	**Stage 1b**	**Stage 2**	**Stage 2a**	**Stage 2b**
Training time	2 weeks	3 days	18 h	10.5 days	3 days	1 day
# GPUs	32	8	8	32	32	24
# samples (steps)	3.76M	376K	26K	1.9M	128K	26K
Learning rate	5e-5	3e-6	3e-6	5e-5	5e-6	3e-6
Learning rate scheduler	cosine	none	none	cosine	none	none
Special LR	2e-4[a]	none	none	2e-4[b]	none	none
Dataset	ERA5	ERA5	analysis	ERA5	ERA5	analysis
Loss type	MAE	MAE	MAE	CRPS	CRPS	CRPS
Auto-regressive steps	1	2	2	1	4	4
# members	1	1	1	2	2	2
Noise injection	No	No	No	Yes	Yes	Yes
Variable lead time	0-12 h	No	No	0-12 h	No	No
Drop path	0.1	0.1	0.0	0.0	0.0	0.0

[a] Applied to encoder and decoder surface variable weights, time embedding weights, and layer normalization weights.
[b] Applied to the backbone's noise embedding MLP weights and layer norm weights.

6.2 Other minor updates in Aurora 1.5

Inclusion of top-of-atmosphere insolation. While out of the scope of this current work, some minor features for Aurora 1.5 were designed with long-term stability of the auto-regressive rollouts in mind. Recent work has shown that Aurora produces impressive and stable long-term forecasts, with the seasonal cycle reproduced thanks to the use of absolute time embeddings [Lehmann et al., 2026]. As such, adding the incoming solar radiation, a pre-computed value that is a function of location and time and includes diurnal and seasonal cycles, may not be necessary. Nevertheless, our variants of Aurora 1.5 were trained with the insolation added as an additional input in the tensor of surface variables (but is not predicted), and the code is updated to compute insolation for insertion at every auto-regressive step.

Clipping during auto-regressive rollouts. Another tweak for stability involves preventing unphysical predictions from the model's forward pass from being used as inputs in the next rollout step. Having only optimized for regression losses, Aurora is not guaranteed to observe physical constraints (such as precipitation being non-negative). Since all training inputs are physical, we avoid requiring the model to correct its own obvious errors in the rollout fine-tuning. Variables which are explicitly clipped include total column water vapor, clipped to a minimum of 0 kg m^{-3}; total, low, medium, and high cloud cover, clipped to a minimum of 0 and maximum of 1; volumetric soil water in layer 1, clipped to a minimum of 0; sea ice cover, clipped to a minimum of 0 and maximum of 1; and snow depth, clipped to a minimum of 0 and a maximum of 10 m (a prescribed value used for large snow/ice sheets in areas like Greenland). Note that output-only variables are not clipped since they are not used as future inputs, and that clipping is not applied by the code to the output of the model, but only when model predictions are fed back in as inputs.

Updated lead-time embeddings. The default lead-time embeddings in Aurora use a Fourier decomposition of lead-time values with a minimum "wavelength" of 1 minute and a maximum of 3 weeks. The small minimum wavelength can be problematic because the periodicity is much shorter than the typical variation of lead times of 1 hour when using the variable-lead-time samples, so, to avoid over-fitting to specific parts of the sine and cosine waves, we change the minimum embedding wavelength to 6 hours.

6.3 Rolling out ensemble forecasts

Aurora 1.5 can be rolled out without any consideration of the fine 1-h lead time. We define each "major" increment as a 6-hour forecast whereby $\hat{\mathbf{x}}_{T+6} = \Theta(\mathbf{x}_T, \mathbf{x}_{T-6}, \tau)$ with $\tau = 6$, where $\mathbf{x}$ is the atmospheric state vector, $\hat{\mathbf{x}}$ is the prediction, and Θ represents the model. Then, the 12-h forecast is generated by feeding in the prediction $\hat{\mathbf{x}}$ at $T + 6$ with the observed $\mathbf{x}$ at T, and so on. The implementation of variable lead time in Aurora 1.5 means that we can use the same inputs to generate each lead time between $\tau = 0$ and $\tau = 6$. As an example, the prediction at hour 9 is given by $\hat{\mathbf{x}}_{T+9} = \Theta(\hat{\mathbf{x}}_{T+6}, \mathbf{x}_T, 3)$.

To generate an ensemble of M members $\{\hat{x}^{(m)}\}_{m=1}^{M}$, we simply perform M independent forward passes of the Aurora 1.5 ENS model, either from the same initial condition (model perturbations only) or different IFS initial

conditions. Members diverge through application of different noise for each forecast and for each major forecast step (6 hours). When sub-stepping is used (i.e., producing predictions at intermediate lead times within each major step), we employ a noise accumulation strategy to ensure temporal coherence across sub-steps. The model maintains a first-in-first-out (FIFO) cache of the n most recent noise tensors, where n equals the number of sub-steps per main step. At each forward call, a fresh noise sample $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is drawn and appended to the cache, evicting the oldest entry (we start with n entries). The effective noise used for conditioning is then computed as

$$\epsilon_{\text{eff}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i,$$

where the sum runs over all cached samples. This running-average scheme introduces auto-correlation between the noise injected at consecutive sub-steps, producing smoother transitions within each main step. Because the cache size equals the number of sub-steps, after one full main step all noise entries have been replaced, so the effective noise between consecutive main steps is independent, matching the training regimen in which each 6-hour step receives independent stochastic perturbations.

6.4 CRPS training objective

For a single scalar predictand y with target value $y^\star$, the Continuous Ranked Probability Score is

$$\mathrm{CRPS}(F, y^\star) = \int_{-\infty}^{\infty} \big(F(y) - \mathbf{1}\{y \geq y^\star\}\big)^2 \, \mathrm{d}y, \tag{1}$$

where F is the cumulative distribution function of the forecast [Gneiting and Raftery, 2007]. For an empirical M-member ensemble, the fair (unbiased) estimator is

$$\widehat{\mathrm{CRPS}} = \frac{1}{M} \sum_{m=1}^{M} \big|\hat{x}^{(m)} - y^\star\big| - \frac{1}{2\,M(M-1)} \sum_{m=1}^{M} \sum_{m'=1}^{M} \big|\hat{x}^{(m)} - \hat{x}^{(m')}\big|. \tag{2}$$

With only $M = 2$ members per training step, Equation (2) reduces to a simple closed form that is cheap to evaluate and yields an unbiased gradient estimate.

The total training loss is the sum of (2) over all predicted variables, vertical levels and grid points, with per-variable weights based on the choices in Aurora [Bodnar et al., 2025].

7 Data and evaluation

7.1 Training data

Stages 1, 1a, 2, and 2a ERA5 reanalysis fields [Hersbach et al., 2020] on a regular 0.25-degree lat–lon grid, with the variable set described in [Bodnar et al., 2025] and the additional single-level variables listed in Appendix A. Additionally, Aurora 1.5 uses a larger set of static variables compared to Aurora, adding such features as orography angle and slope, vegetation cover, and soil and vegetation types. These static fields are provided with the code and model checkpoints.

Stages 1b and 2b Operational ECMWF HRES analyses from 2018–2023, with the same configuration as the ERA5 data.

7.2 Implementation of the tropical cyclone track metrics

We evaluate tropical cyclone track forecasts from Aurora, Aurora 1.5, Aurora 1.5 ENS (32 members), and ECMWF IFS (deterministic and 50-member ensemble) against observed best-track positions from IBTrACS v04r01 [Knapp et al., 2010].

Evaluation grid. We define a set of storm-initialization-time pairs using a curated grid of tropical cyclones active during 2024–2025, restricted to synoptic initialization times (00, 06, 12, 18 UTC) with at least 24 hours of subsequent IBTrACS coverage. Basins are assigned using the IBTrACS 'BASIN' field, with the South Indian (SI) basin split at 90°E into "Indian" (west) and "Australian" (east, merged with South Pacific). All model comparisons are performed on the intersection of available forecast keys across models to ensure matched samples.

Track error. Forecast position error is computed as the great-circle (haversine) distance between the predicted and observed latitude–longitude at each valid time step. Only forecast times falling within each storm's IBTrACS observation window are retained. We report mean absolute error (MAE) in kilometers, stratified by lead time (days 1–5, evaluated at 30, 54, 78, 102, and 126 hours) and basin.

Ensemble aggregation. Ensemble forecasts are aggregated to mean and median tracks. The ensemble mean longitude is computed via circular (atan2) averaging to handle the antimeridian correctly; the median longitude uses circular unwrapping relative to the first member. Before aggregation, individual ensemble members deviating more than 15° in latitude or longitude from the per-timestep median are flagged as outliers.

Statistical significance. Relative track errors between models are assessed using a stratified bootstrap procedure. For each of $N = 1{,}000$ iterations, storm identifiers (SIDs) are resampled with replacement within each basin, and the relative error $(\mathrm{MAE}_{\mathrm{model}} - \mathrm{MAE}_{\mathrm{ref}})/\mathrm{MAE}_{\mathrm{ref}}$ is recomputed. A difference is deemed statistically significant if the 95% bootstrap confidence interval excludes zero.

Ensemble calibration. We assess ensemble dispersion using rank histograms. At each forecast time step, ensemble member positions are ranked by great-circle distance from the ensemble centroid; the observed position is then ranked among the members to yield a normalized rank in $(0, 1)$. A well-calibrated ensemble produces a uniform rank distribution. To enable fair comparison between ensembles of different sizes, we also compute subsampled rank histograms by randomly drawing 32 of 50 ECMWF members, averaged over 1,000 repetitions.

A Single-level variable list

All single-level variables used by Aurora 1.5, sourced from ERA5 [Hersbach et al., 2020], are listed in Table 3. For the three radiation variables and the two precipitation variables (total precipitation and snowfall), Aurora 1.5 predicts the accumulation over a one-hour period ending at the valid time.

Table 3: Single-level variables in Aurora 1.5. A ✓ in *In Aurora* indicates the variable was present in the pretrained Aurora model. A ✓ in *Output only* indicates the variable is predicted by the decoder but not provided as encoder input.

Short name	Variable name	Units	In Aurora	Output only
`2t`	2 m temperature	K	✓	
`10u`	10 m u-component of wind	$\mathrm{m\,s^{-1}}$	✓	
`10v`	10 m v-component of wind	$\mathrm{m\,s^{-1}}$	✓	
`msl`	Mean sea-level pressure	Pa	✓	
`2d`	2 m dewpoint temperature	K		
`tcwv`	Total column water vapour	$\mathrm{kg\,m^{-2}}$		
`tcc`	Total cloud cover	(0–1)		
`100u`	100 m u-component of wind	$\mathrm{m\,s^{-1}}$		
`100v`	100 m v-component of wind	$\mathrm{m\,s^{-1}}$		
`sp`	Surface pressure	Pa		
`lcc`	Low cloud cover	(0–1)		
`mcc`	Medium cloud cover	(0–1)		
`hcc`	High cloud cover	(0–1)		
`skt`	Skin temperature	K		
`stl1`	Soil temperature, level 1	K		
`swvl1`	Volumetric soil water, layer 1	$\mathrm{m^3\,m^{-3}}$		
`siconc`	Sea-ice cover	(0–1)		
`sd`	Scaled snow depth	–		
`i10fg`	Instantaneous 10 m wind gust	$\mathrm{m\,s^{-1}}$		✓
`blh`	Boundary-layer height	m		✓
`uvb`	Downward UV radiation at the surface	$\mathrm{J\,m^{-2}}$		✓
`ssrd`	Surface solar radiation downwards	$\mathrm{J\,m^{-2}}$		✓
`ttr`	Top net thermal radiation	$\mathrm{J\,m^{-2}}$		✓
`tp`	Total precipitation (scaled)	m		✓
`sf`	Snowfall (scaled)	m		✓

B Additional figures

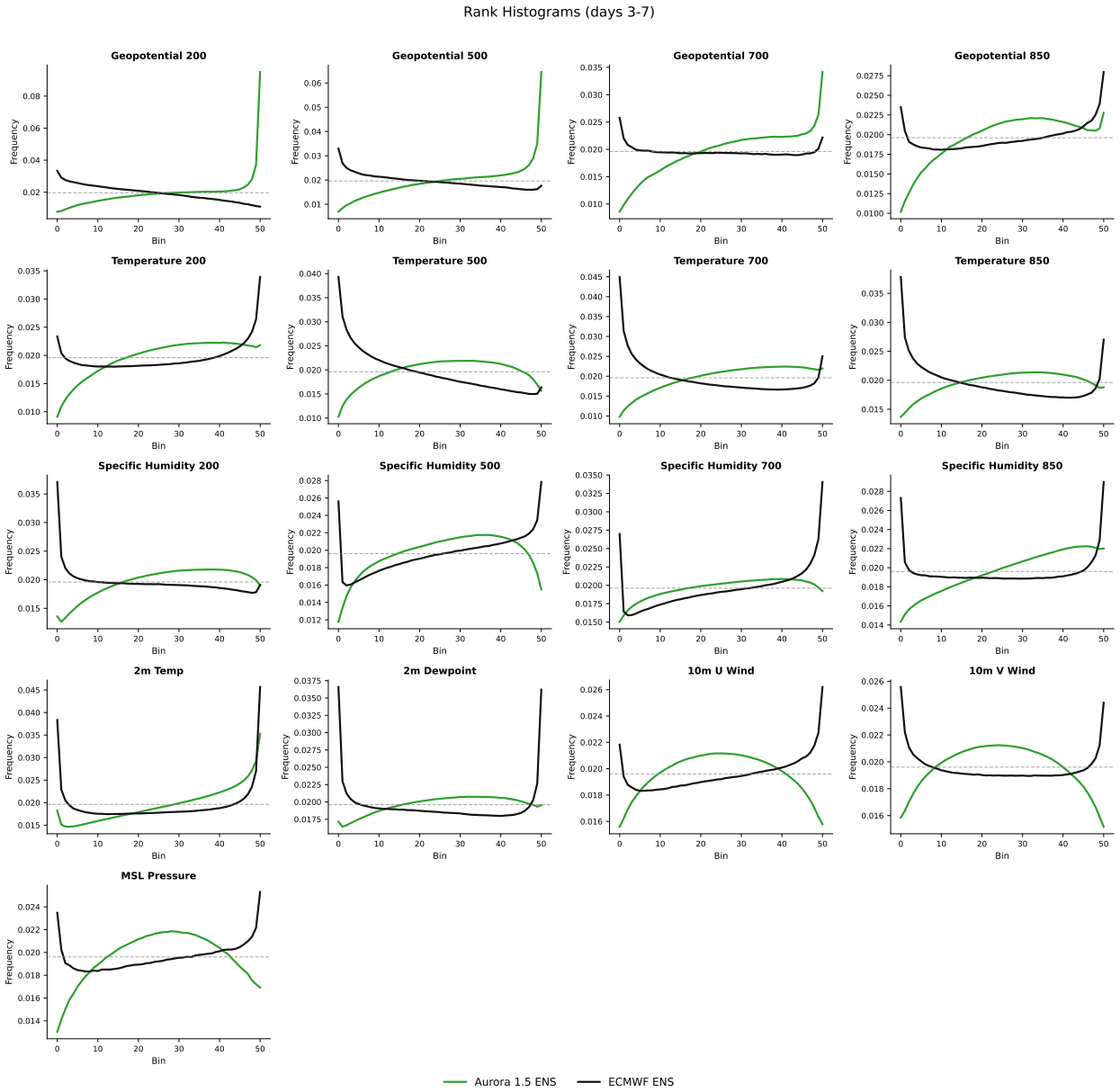

Figure 7: **Rank histograms for Aurora 1.5 ENS and ECMWF ENS** across a range of predicted variables. Values indicate the frequency of occurrence of the truth value falling at a specific rank in the ordered ensemble. Typically, an ensemble is under-dispersive (i.e., it has too little spread) if the histogram is U-shaped, indicating that many true values fall outside the ensemble's predicted distribution. Forecasts for lead days 3–7 are averaged.

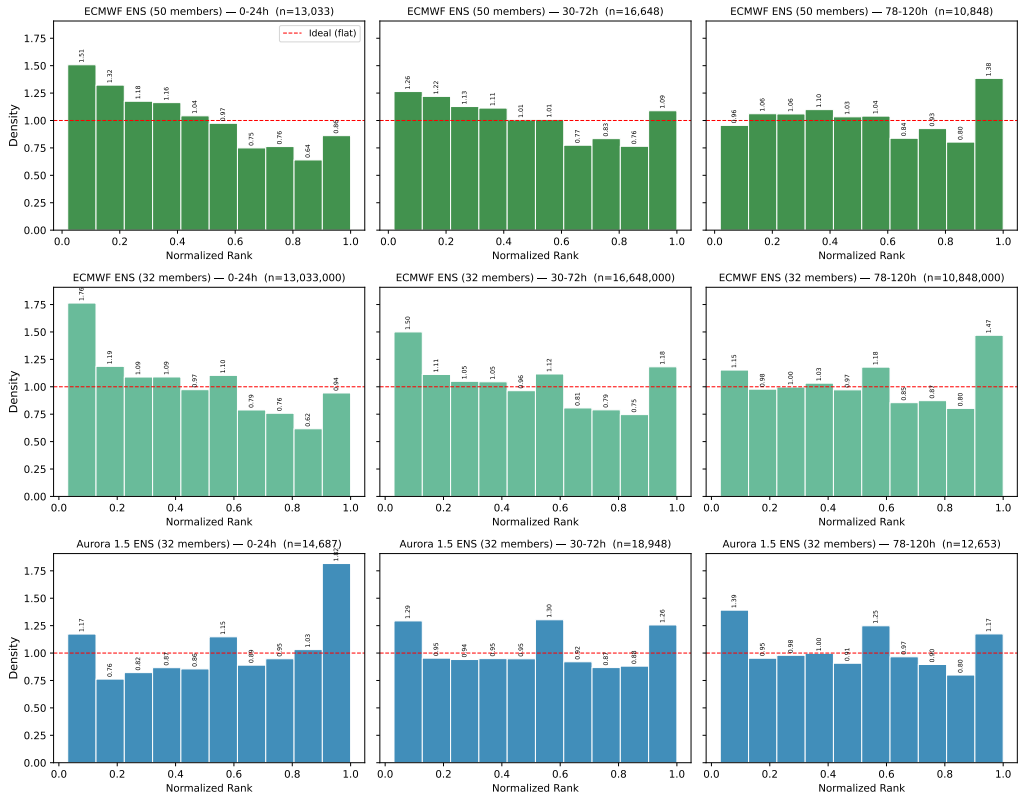

Figure 8: **Rank histograms of tropical cyclone track locations** for ECMWF (top), ECMWF subsampled to 32 members (middle), and Aurora 1.5 ENS. For each forecast time, we compute the ensemble centroid and rank the verifying IBTrACS position by its distance from that centroid relative to the member distances. Middle ranks indicate that the observed storm lies at a typical member radius; ranks near 1 indicate that the observation lies further from the centroid than all members, indicating under-dispersion. A flat histogram would be ideal under this radial-rank diagnostic.