Calibration of Large Neural Weather Models

Andre Graubner
NVIDIA Corporation
Zurich, Switzerland

Kamyar Azizzadenesheli
NVIDIA Corporation
Santa Clara, CA 95051

Jaideep Pathak
NVIDIA Corporation
Santa Clara, CA 95051

Morteza Mardani
NVIDIA Corporation
Santa Clara, CA 95051

Mike Pritchard
NVIDIA Corporation
Santa Clara, CA 95051

Karthik Kashinath
NVIDIA Corporation
Santa Clara, CA 95051

Animashree Anandkumar
California Institute of Technology
Pasadena, CA 91125
NVIDIA Corporation
Santa Clara, CA 95051

Abstract

Uncertainty quantification of weather forecasts is a necessity for reliably planning for and responding to extreme weather events in a warming world. This motivates the need for well-calibrated ensembles in probabilistic weather forecasting. We present initial results for the calibration of large-scale deep neural weather models for data-driven probabilistic weather forecasting. By explicitly accounting for uncertainties about the forecast's initial condition and model parameters, we generate ensemble forecasts that show promising results on standard diagnostics for probabilistic forecasts. Specifically, we are approaching the Integrated Forecasting System (IFS), the gold standard on probabilistic weather forecasting, on: (i) the spread-error agreement; and (ii) the Continuous Ranked Probability Score (CRPS). Our approach scales to state-of-the-art data-driven weather models, enabling cheap post-hoc calibration of pretrained models with tens of millions of parameters and paving the way towards the next generation of well-calibrated data-driven weather models.

1 Introduction

Data-driven neural weather models [Sønderby et al., 2020] [Pathak et al., 2022] [Hu et al., 2022] promise to revolutionize our capacity to predict extreme weather events, mitigate their disastrous impacts, manage energy systems, and democratize access to high-quality weather forecasts. In contrast to traditional, hand-crafted numerical weather prediction (NWP) systems like the IFS by the European Centre for Medium-Range Weather Forecasts (ECMWF), these deep learning-based systems learn to forecast future weather conditions given an initial condition from vast amounts of historical observational or reanalysis data. This enables orders-of-magnitude speedups over state-of-the-art NWP systems on modern hardware accelerators, while potentially matching or surpassing NWP forecasting skill [Schultz et al., 2021].

Earth's weather is a highly complex, multi-scale, nonlinear, chaotic system. Weather forecasts are made based on incomplete knowledge of initial conditions and uncertainties in the model's characterization of the physics of the Earth system. Hence, it is crucial to quantify uncertainties in our predictions to enable informed decision making. Therefore, a requirement for most applications

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

of weather prediction systems are probabilistic forecasts. These probabilistic forecasts often take the form of ensemble forecasts [Houtekamer and Derome, 1995] [Atger, 1999], predicting n different trajectories using perturbed initial conditions or perturbed parameters in models to statistically quantify forecast uncertainties. Deep neural networks, however, just like deterministic NWP models, are known to be overconfident in their predictions, unless careful countermeasures are taken [Nguyen et al., 2015] [Guo et al., 2017].

In this paper we show that model-agnostic perturbation strategies based on spatially correlated scale-aware initial condition noise and Bayesian model uncertainty leads to well-calibrated ensembles of trajectories and consistent calibration gains over the baseline, measured by metrics routinely used to quantify reliability of probabilistic weather forecasts: (i) spread-error agreement; and (ii) the CRPS.

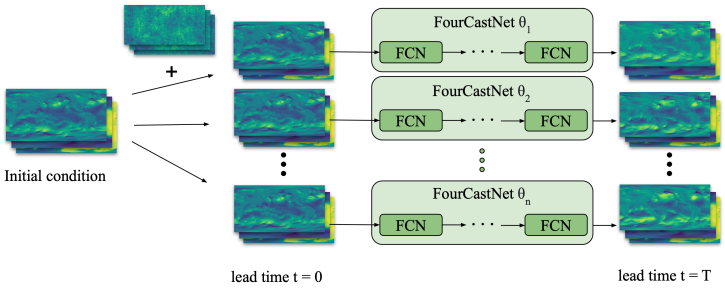

Figure 1: The ensembling strategy presented in this paper: (i) Perturbing initial conditions with spatially correlated noise to produce n initial states; and (ii) Sampling n different FourCastNet models from a posterior distribution over models. Each model auto-regressively predicts a unique trajectory.

2 Methods

2.1 FourCastNet and ERA5 Reanalysis

We base this work on the state-of-the-art FourCastNet [Pathak et al., 2022], a data-driven high-resolution global weather model. FourCastNet is an auto-regressive transformer-based neural network based on the Adaptive Fourier Neural Operator [Guibas et al., 2021]. Given an initial condition, FourCastNet predicts the atmospheric state 6 hours into the future. The output of the network then gets fed back into the network to iteratively produce subsequent forecasts, time-step by time-step, up to two weeks. FourCastNet is highly scalable [Kurth et al., 2022] and its skill has been shown to be approaching ECMWF's IFS, including on extreme weather events such as hurricanes, tropical cyclones, and extratropical cyclones.

Here, we take the ground-truth for training and evaluation to be a 26-variable subset of the ERA5 global reanalysis dataset [Hersbach et al., 2020] provided by ECMWF. The input and output of the neural network are therefore 26 meteorological variables at a global spatial resolution of 0.25 degrees, or 25 km (represented by a tensor of shape [26, 720, 1440]).

2.2 Initial condition uncertainty

To create ensemble forecasts we perturb the initial condition (IC), i.e. the weather conditions from which we start the forecast. We thus run n trajectories from perturbed versions of our initial state.

We empirically notice that perturbing the initial condition by uncorrelated Gaussian noise leads to under-dispersive ensembles: The noise dissipates quickly and does not produce an appropriate amount of spread. This is our baseline.

We hypothesize that if FourCastNet acts like a real dynamical system, we should expect the network to have physically plausible sensitivity to the initial condition. Thus, the spatial scale of perturbations should matter. To test this hypothesis, we add spatial correlation to the perturbations. We use pink noise with a power spectral density inversely proportional to the sum of the x and y frequencies of the 2-dimensional noise signal ($\frac{1}{f_x + f_y}$) .

2.3 Model uncertainty

Another source of forecast uncertainty is that of the forecast model itself. Since the model is trained on a finite amount of data, and the reanalysis dataset uses a model with incomplete and unknown physics, uncertainty about the parameters of the neural network remains even after training. A popular approach to uncertainty quantification in deep learning is to explicitly account for this so-called epistemic uncertainty using Bayesian methods [Wilson and Izmailov, 2020].

To improve calibration of FourCastNet, we employ an efficient post-hoc approximation of $p(\theta|D)$, the posterior distribution over model weights θ given training data D. By sampling different models θ_i during inference, we achieve diverse ensemble forecasts.

As demonstrated by Maddox et al. [2019], we run stochastic gradient descent with a high, constant learning rate starting from the fully trained FourCastNet model θ_{MAP} and take $\theta_{\text{SWA}} = \overline{\theta_t} = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ to be the mean of the different weight vectors along the SGD trajectory. We can approximate the true posterior as a multivariate Gaussian distribution, where, to make the approximation tractable, we decompose the covariance of this distribution as the combination of a diagonal matrix $\Sigma_{\text{diag}} = diag(\overline{\theta_t^2} - \overline{\theta_t}^2)$ and a low-rank matrix $\Sigma_{\text{low-rank}} = \frac{1}{K-1}\cdot\widehat{D}\widehat{D}^T$, where $\widehat{D}$ is a matrix with columns corresponding to the deviations of the last K checkpoints from the sample mean θ_{SWA}. Intuitively, the diagonal matrix captures the uncertainty for each weight individually, while the low-rank matrix corresponds to the sample-covariance of the SGD iterates of the last K checkpoints, where K controls the rank of covariance matrix: Since $\widehat{D}$ has only K columns, $\Sigma_{\text{low-rank}}$ has at most rank K.

Following Maddox et al. [2019], we now approximate the posterior as $\mathcal{N}(\theta_{\text{SWA}}, \frac{1}{2}\cdot(\Sigma_{\text{diag}} + \Sigma_{\text{low-rank}}))$.

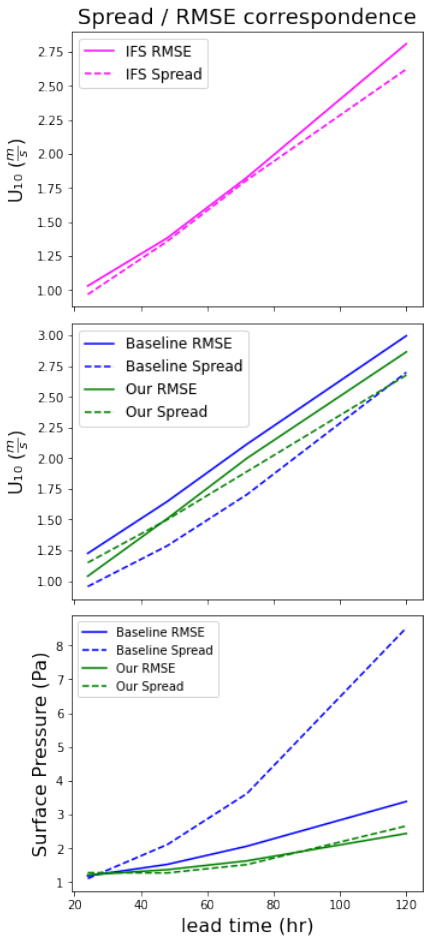

Figure 2: Ensemble spreads (square root of average ensemble variance) and skills (RMSE) as a function of lead time for U_{10} (the 10m U-component of wind speed) and surface pressure

3 Results

Evaluating the quality of probabilistic weather forecasting systems is a multi-dimensional problem, including many qualitative and quantitative metrics and diagnostics [Murphy, 1993, Christensen et al., 2014]. Since the specific requirements of any real forecasting system depend on the intended use-cases of the system, we focus on general-purpose diagnostics here that are applicable to global forecast ensembles, and leave a thorough use-case-specific investigation for future work.

3.1 Spread-Error Agreement

A well-calibrated ensemble should have its forecast spread match its forecast error. Following Fortin et al. [2014], we define the ensemble spread as the square root of the average ensemble variance, and verify that the forecast skill (RMSE of the ensemble mean) approximately equals this spread.

Figure 2 (top panel) illustrates this using a 50-member TIGGE [Bougeault et al., 2010] IFS ensemble. Here, spread and RMSE are approximately equal. We run 50-member ensemble predictions for 10 non-overlapping time-frames in the out-of-sample year 2018 for our baseline (uncorrelated Gaussian noise, no Bayesian model) and our Bayesian model with pink noise.

Figure 2 (middle and bottom panels) show the spread-error agreement for different variables. Note that the Bayesian model with correlated IC perturbations (pink noise) has higher skill (lower RMSE),

with a spread profile that matches the RMSE more closely than the baseline. Also note that while the baseline ensemble is generally under-dispersive (spread lower than RMSE), for some variables like surface pressure, the ensemble spread diverges. This could be alleviated by reducing the magnitude of IC perturbations, however, at the cost of even stronger under-dispersiveness on other variables.

3.1.1 CRPS

A common metric for probabilistic forecasts of continuous variables is the CRPS [Matheson and Winkler, 1976, Leu, 2020]. It scores a predicted distribution against the actually observed outcome and encourages well-calibrated predictions. If F denotes the CDF of a predicted distribution and y the observed outcome, then we define $\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(y') - \mathbb{1}(y' - y))^2 \, dy'$, where $\mathbb{1}$ denotes the heavy-side step function. This metric is useful, since we generally do not have access to the actual probability distribution over future weather, but only realisations sampled from this distribution. In order to minimize the expected CRPS, we need our predicted distribution to be equal to the actual (unknown) distribution over future weather states.

In our case, we obtain the forecast CRPS by averaging the pixel-wise CRPS over the entire domain. For comparison, we contrast our 50-member ensemble results with the CRPS of a 50-member TIGGE [Bougeault et al., 2010] IFS ensemble.

Encouraging improvements are found in all forecasted variables with CRPS reductions beginning to approach the IFS standard. This is reassuring especially given the simplicity of our uncertainty quantification approach. We expect that refinements to this approach, such as incorporating multivariate spatio-temporal correlations in IC noise that is injected, could yield additional gains in calibration.

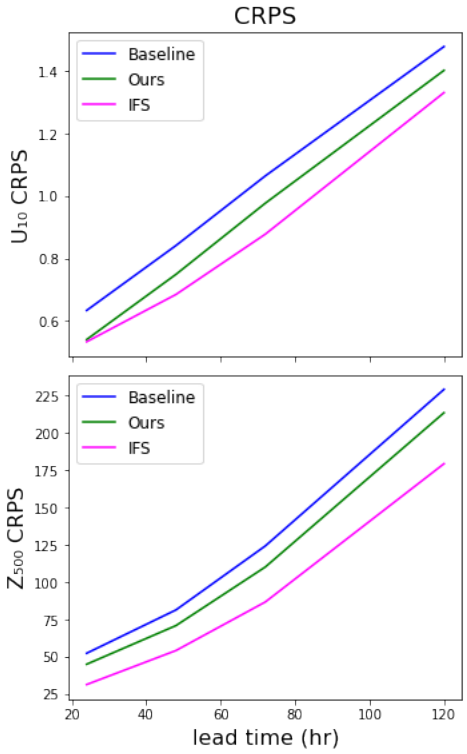

Figure 3: Ensemble CRPS comparing the baseline FourCastNet, the Bayesian FourCastNet and the IFS ensemble

4 Conclusion and future work

We have presented initial evidence that large neural weather models can be used to produce well-calibrated ensemble forecasts using scale-aware spatially correlated initial condition noise and Bayesian deep learning methods for quantifying model uncertainty. While these methods do not surpass state-of-the-art NWP models in terms of standard calibration metrics like CRPS, we view these first successes as a promising sign: The sensitivity of FourCastNet to spatially correlated noise serves as evidence that FourCastNet acts like Earth's chaotic weather system.

In future work we plan to exploit correlation between physical variables and between time-steps for more principled perturbations. In addition to these perturbed initial conditions, more sophisticated approaches to Bayesian model uncertainty are possible, and might indeed be required to gracefully scale to much larger models. While SWA-G can scale to large models, recent results [Daxberger et al., 2021] suggest that we might be able to outperform SWA-G at lower cost by performing more expressive posterior approximation on a subset of our network.

References

Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation, author=Martin Leutbecher and Thomas Haiden. *Quarterly Journal of the Royal Meteorological Society*, 147-734:425–442, 2020.

Frederic Atger. The skill of ensemble prediction systems. *Monthly Weather Review*, 127(9):1941–1953, 1999.

Philippe Bougeault, Zoltan Toth, Craig Bishop, Barbara Brown, David Burridge, Hui De Chen, Beth Ebert, Manuel Fuentes, Thomas M. Hamill, Ken Mylne, Jean Nicolau, Tiziana Paccagnella, Young Youn Park, David Parsons, Baudouin Raoult, Doug Schuster, Pedro Silva Dias, Richard Swinbank, Yoshiaki Takeuchi, Warren Tennant, Laurence Wilson, and Steve Worley. The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91: 1059–1072, 8 2010. ISSN 00030007. doi: 10.1175/2010BAMS2853.1.

H. M. Christensen, I. M. Moroz, and T. N. Palmer. Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141-687:538–549, 2014.

Erik A. Daxberger, Eric T. Nalisnick, James Urquhart Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian Deep Learning via Subnetwork Inference. In *ICML*, 2021.

Vincent Fortin, M. Abaza, F. Anctil, and R. Turcotte. Why should ensemble spread match the RMSE of the ensemble mean? *Journal of Hydrometeorology*, 15:1708–1713, 8 2014. ISSN 15257541. doi: 10.1175/JHM-D-14-0008.1.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier Neural Operators: Efficient Token Mixers for Transformers. 11 2021. URL `http://arxiv.org/abs/2111.13587`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

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*, 146:1999–2049, 7 2020. ISSN 1477870X. doi: 10.1002/qj.3803.

PL Houtekamer and Jacques Derome. Methods for ensemble prediction. *Monthly Weather Review*, 123(7):2181–2196, 1995.

Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. SwinVRNN: A Data-Driven Ensemble Forecasting Model via Learned Distribution Perturbation. 5 2022. URL `http://arxiv.org/abs/2205.13158`.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: Accelerating Global High-Resolution Weather Forecasting using Adaptive Fourier Neural Operators. *arXiv preprint arXiv:2208.05419*, 2022.

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. 2 2019. URL `http://arxiv.org/abs/1902.02476`.

James E. Matheson and Robert L. Winkler. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22:1087–1096, 1976.

Allan H. Murphy. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, 8:281–293, 1993.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. 2 2022. URL `http://arxiv.org/abs/2202.11214`.

M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler. Can deep learning beat numerical weather prediction?, 4 2021. ISSN 1364503X.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A Neural Weather Model for Precipitation Forecasting. 3 2020. URL `http://arxiv.org/abs/2003.12140`.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.