ANISOTROPIC SPECTRAL ERROR DRESSING FOR CALIBRATED ENSEMBLE WEATHER FORECASTS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Data-driven weather models achieve remarkable deterministic skill but lack native uncertainty quantification. Existing post-processing methods that convert deterministic forecasts into probabilistic ensembles typically assume isotropic error structure, ignoring directional patterns in forecast errors. We show that GraphCast forecast errors exhibit significant quasi-zonal anisotropy, with zonal modes containing $4.26\times$ more power than meridional modes. To exploit this structure, we propose Anisotropic Spectral Error Dressing (ASED), a training-free method that models within-degree anisotropy via the normalized order ratio $\mu = |m|/l$, partitioning modes into 3 μ-bins across 4 degree bands. On WeatherBench2 Z500 at 5-day lead time, ASED achieves 2.92% global CRPS improvement over standard spectral error dressing, with 82.4% of gridpoints showing improvement. Our results demonstrate that exploiting directional error structure can meaningfully improve probabilistic calibration without model retraining.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Machine learning has transformed global weather forecasting. Models such as GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), and FourCastNet (Pathak et al., 2022) now achieve deterministic skill competitive with operational numerical weather prediction (NWP) systems, while requiring orders of magnitude less computation at inference time. These advances have been systematically evaluated on benchmarks like WeatherBench (Rasp et al., 2020) and WeatherBench2 (Rasp et al., 2023), establishing data-driven weather forecasting as a viable alternative to physics-based models.

However, most AI weather models produce only point forecasts, lacking the uncertainty quantification essential for downstream decision-making in energy trading, logistics, and emergency planning. Operational NWP centers address this through ensemble prediction systems that run multiple perturbed forecasts, but generating such ensembles from deterministic AI models remains challenging. Recent approaches include training generative models like GenCast (Price et al., 2024) or fine-tuning with perturbed initial conditions (Zhong et al., 2024), but these require substantial computational resources or model retraining.

A simpler alternative is post-processing: adding structured perturbations to deterministic forecasts based on historical error statistics. Spectral Error Dressing (SED) samples perturbations in spherical harmonic space with variance matching the empirical degree spectrum C_l, but assumes isotropy within each degree—all orders m have equal expected power. This assumption may be violated in practice: atmospheric forecast errors often exhibit directional structure, particularly in regions dominated by zonal flow patterns.

We hypothesize that AI weather forecast errors are significantly anisotropic, with quasi-zonal modes containing more power than meridional modes, and that exploiting this structure can improve probabilistic calibration. To test this, we propose **Anisotropic Spectral Error Dressing (ASED)**, which extends SED by modeling within-degree anisotropy via the normalized order ratio $\mu = |m|/l$.

[1] https://gitlab.com/fars-a/anisotropic-spectral-error-dressing-weatherbench2

ASED partitions modes into 3 μ-bins across 4 degree bands, redistributing variance while preserving the degree spectrum.

Our contributions are threefold. First, we demonstrate that GraphCast Z500 forecast errors exhibit significant quasi-zonal anisotropy ($A_{\text{cal}} = -0.276$), with zonal modes containing $4.26\times$ more power than meridional modes. Second, we propose ASED, a training-free post-processing method that models within-degree anisotropy via μ-binning and multi-band degree decomposition. Third, we achieve 2.92% global CRPS improvement over SED on WeatherBench2, with 82.4% of grid-points showing improvement, demonstrating that exploiting directional error structure meaningfully improves probabilistic calibration.

2 Related Work

Data-Driven Weather Models. Recent advances in machine learning have produced global weather models that rival or exceed traditional numerical weather prediction (NWP) in deterministic skill. GraphCast (Lam et al., 2023) employs graph neural networks to achieve state-of-the-art forecasts at 0.25° resolution, while Pangu-Weather (Bi et al., 2023) uses 3D vision transformers for efficient multi-level prediction. FourCastNet (Pathak et al., 2022) leverages adaptive Fourier neural operators for high-resolution forecasting, and FuXi (Chen et al., 2023) introduces a cascade architecture for extended-range prediction. These models are evaluated on standardized benchmarks such as WeatherBench (Rasp et al., 2020) and WeatherBench2 (Rasp et al., 2023). However, these deterministic models lack native uncertainty quantification.

Ensemble Generation for AI Weather. Several approaches address the uncertainty gap in AI weather models. GenCast (Price et al., 2024) uses diffusion models to generate probabilistic forecasts directly, while FuXi-ENS (Zhong et al., 2024) fine-tunes deterministic models with perturbed initial conditions. Mahesh et al. (2024) propose Huge Ensembles using spherical Fourier neural operators for efficient large-ensemble generation. Bulte et al. (2024) provide a comprehensive survey of uncertainty quantification methods for data-driven weather models. These approaches typically require substantial computational resources or model retraining.

Statistical Post-Processing. Classical post-processing methods calibrate ensemble forecasts using statistical techniques. Ensemble Model Output Statistics (EMOS) fits parametric distributions to ensemble output (Feldmann et al., 2015), while Ensemble Copula Coupling (ECC) preserves spatial correlation structure (Schefzik et al., 2013; Bouallègue et al., 2016). Neural network approaches extend these methods with learned calibration functions (Rasp & Lerch, 2018). Conformal prediction offers distribution-free uncertainty quantification (Gopakumar et al., 2024). Our work differs by operating in spectral space to exploit error structure.

Spectral Methods for Atmospheric Fields. Spherical harmonic analysis provides a natural basis for global atmospheric fields. Lang & Schwab (2015) develop theory for isotropic Gaussian random fields on the sphere, while Cao et al. (2022) extend to locally anisotropic covariance functions. Subich et al. (2025) use spherical harmonic loss functions to address the double penalty problem in AI weather training. Our ASED method builds on spectral error dressing by introducing within-degree anisotropy modeling via the $\mu = |m|/l$ parameterization.

3 Method

We present Anisotropic Spectral Error Dressing (ASED), a training-free method for converting deterministic AI weather forecasts into calibrated probabilistic ensembles. ASED extends standard spectral error dressing by modeling within-degree anisotropy in forecast errors.

3.1 Problem Setup

Let $f(t, x)$ denote a deterministic forecast at initialization time t and grid point x, and let $a(t+\tau, x)$ denote the verifying analysis (ERA5) at lead time τ. The forecast residual is $r(t, x) = f(t, x) - a(t+\tau, x)$. Our goal is to generate an M-member ensemble $\{f^{(j)}(t, x)\}_{j=1}^{M}$ with calibrated uncertainty,

where each member is formed by adding a structured perturbation: $f^{(j)} = f_{bc} + \eta^{(j)}$, with f_{bc} being a bias-corrected forecast.

3.2 Spherical Harmonic Representation

Global atmospheric fields are naturally represented using spherical harmonics $Y_{lm}(\theta, \phi)$, which form an orthonormal basis on the sphere. Any square-integrable field can be expanded as:

$$r(\theta, \phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} r_{lm} Y_{lm}(\theta, \phi), \tag{1}$$

where l is the degree (related to spatial scale) and m is the order (related to zonal wavenumber). The degree power spectrum characterizes scale-dependent variance:

$$C_l = \frac{1}{2l+1} \sum_{m=-l}^{l} |r_{lm}|^2. \tag{2}$$

Standard Spectral Error Dressing (SED) samples perturbations η with variance matching C_l, treating all orders m within each degree l as having equal expected power. This corresponds to an isotropic Gaussian random field on the sphere (Lang & Schwab, 2015). However, atmospheric forecast errors often exhibit directional structure, particularly in the extra-tropics where errors organize along mid-latitude jets and storm tracks.

3.3 Within-Degree Anisotropy

To characterize directional structure within each degree, we introduce the normalized order ratio:

$$\mu = \frac{|m|}{l} \in [0, 1]. \tag{3}$$

For large l, this ratio approximates the orientation of a mode's wavevector: $\mu \to 0$ corresponds to quasi-zonal modes (large meridional gradients, slow longitudinal variation), while $\mu \to 1$ corresponds to quasi-meridional modes. This interpretation follows from the plane-wave analogy where $k_{\text{zonal}} \approx |m|$ and $k_{\text{meridional}} \approx \sqrt{l(l+1) - m^2}$ (Subich et al., 2025).

We quantify anisotropy using a calibration index computed on historical residuals. Partitioning modes into low-μ ($\mu < 0.5$) and high-μ ($\mu \geq 0.5$) bins, we compute the average power in each bin:

$$P_{\text{low}}(l) = \mathbb{E}_t \left[\text{mean}_{m:|m|/l<0.5} |r_{lm}(t)|^2 \right], \quad P_{\text{high}}(l) = \mathbb{E}_t \left[\text{mean}_{m:|m|/l\geq 0.5} |r_{lm}(t)|^2 \right]. \tag{4}$$

The anisotropy index is then:

$$A_{\text{cal}} = \mathbb{E}_{l \geq 10} \left[\frac{P_{\text{high}}(l) - P_{\text{low}}(l)}{P_{\text{high}}(l) + P_{\text{low}}(l)} \right]. \tag{5}$$

Negative A_{cal} indicates quasi-zonal bias (more power in low-μ modes), while $|A_{\text{cal}}| < 0.1$ suggests approximate isotropy.

3.4 ASED Formulation

ASED extends SED by redistributing variance across orders m while preserving the degree spectrum C_l. We partition modes into $K = 3$ μ-bins (zonal: $\mu < 0.33$, intermediate: $0.33 \leq \mu < 0.67$, meridional: $\mu \geq 0.67$) and degrees into $B = 4$ bands (planetary: $l \in [10, 23]$, synoptic: $l \in [24, 59]$, mesoscale: $l \in [60, 146]$, small-scale: $l \in [147, L]$).

For each (μ-bin, band) combination, we estimate weights $w_{\mu,b}$ from calibration residuals. The perturbation generation proceeds as follows:

Step 1: Sample base coefficients. Draw i.i.d. Gaussian coefficients $\epsilon_{lm} \sim \mathcal{N}(0, 1)$.

Step 2: Apply anisotropic scaling. Define per-mode variance multipliers:

$$g_{lm} = \frac{w_{\text{bin}(|m|/l),\text{band}(l)}}{\bar{w}_l}, \quad \text{where} \quad \bar{w}_l = \frac{1}{2l+1} \sum_{m=-l}^{l} w_{\text{bin}(|m|/l),\text{band}(l)}. \tag{6}$$

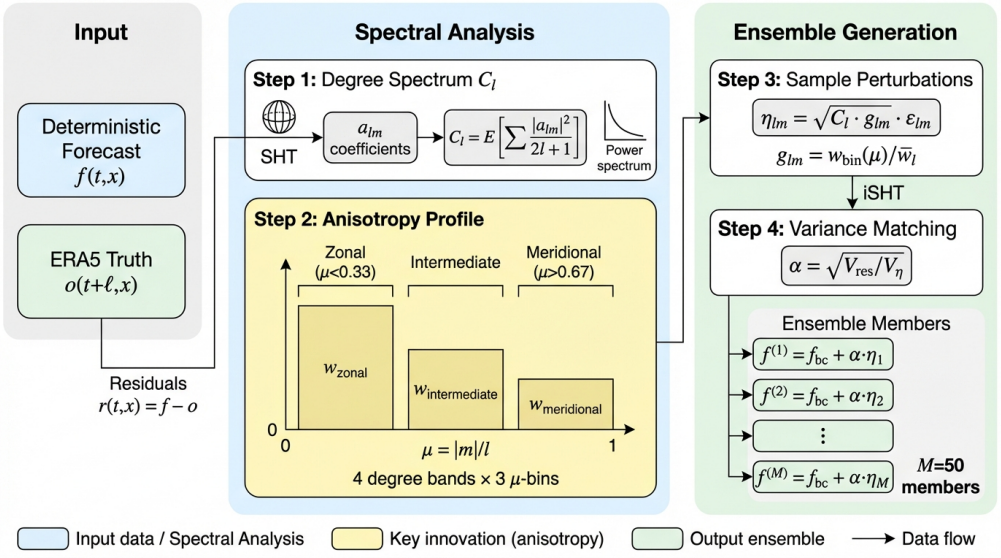

Figure 1: Overview of Anisotropic Spectral Error Dressing (ASED). Given a deterministic GraphCast forecast, ASED: (1) computes spherical harmonic coefficients of historical residuals, (2) estimates degree spectrum C_l and within-degree anisotropy weights w_μ across 3 μ-bins and 4 degree bands, (3) generates $M = 50$ ensemble members by sampling Gaussian perturbations with the calibrated anisotropic spectral structure.

The normalization by $\bar{w}_l$ ensures that the mean variance within each degree equals C_l.

Step 3: Form perturbation coefficients.

$$\eta_{lm} = \sqrt{C_l \cdot g_{lm}} \cdot \epsilon_{lm}. \tag{7}$$

Step 4: Transform to physical space. Apply the inverse spherical harmonic transform: $\eta(x) = \text{iSHT}(\eta_{lm})$.

Step 5: Global variance matching. Scale perturbations by $\alpha = \sqrt{V_{\text{res}}/V_\eta}$, where V_{res} and V_η are the spatial variances of residuals and perturbations, respectively.

The final ensemble members are $f^{(j)}(t,x) = f_{bc}(t,x) + \alpha \cdot \eta^{(j)}(x)$.

Figure 1 illustrates the ASED pipeline. The key innovation is the μ-binning mechanism that redistributes variance across orders while preserving the degree spectrum, enabling the method to capture directional error structure without additional parameters beyond the bin weights.

4 Experiments

We evaluate ASED on the WeatherBench2 benchmark (Rasp et al., 2023), comparing against spectral and non-spectral baselines for probabilistic calibration of deterministic AI weather forecasts.

4.1 Experimental Setup

Dataset and Task. We use GraphCast (Lam et al., 2023) forecasts from WeatherBench2 for year 2020, focusing on 500 hPa geopotential height (Z500) at 5-day lead time. Z500 is a standard tracer of mid-latitude flow and is commonly used for evaluating synoptic-scale forecast skill. We use a time-based split: odd months (Jan, Mar, May, Jul, Sep, Nov) for calibration and even months (Feb, Apr, Jun, Aug, Oct, Dec) for evaluation, yielding 354 evaluation forecasts.

Baselines. We compare against four methods: (1) **Deterministic**: bias-corrected GraphCast point forecast (ensemble size $M = 1$); (2) **Isotropic GP**: Gaussian process perturbations with 1200 km

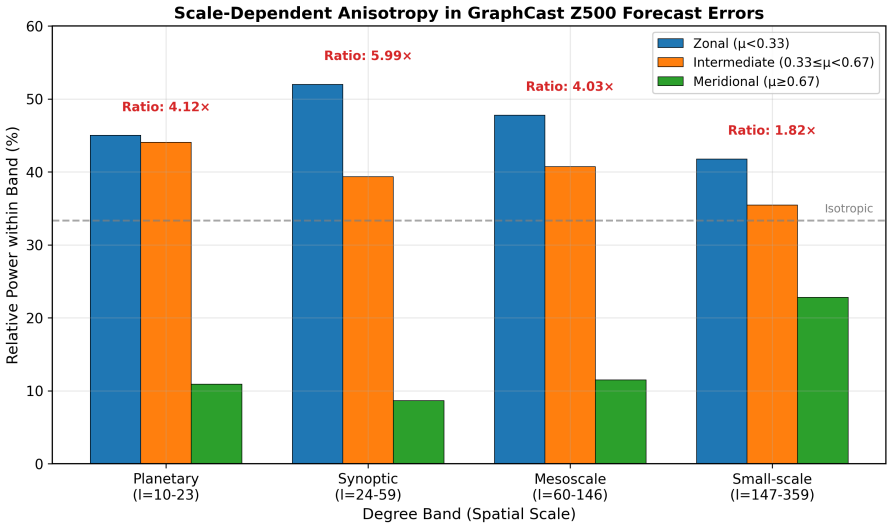

Figure 2: Scale-dependent anisotropy in GraphCast Z500 forecast errors. Each bar group shows the relative power in 3 μ-bins (zonal: $\mu < 0.33$, intermediate: $0.33 \leq \mu < 0.67$, meridional: $\mu \geq 0.67$) for 4 degree bands. Synoptic scales ($l = 24$–59) show the strongest anisotropy with zonal/meridional ratio of 5.99. Dashed line indicates isotropic reference (33.3 %).

length scale, matching the baseline used in GenCast (Price et al., 2024); (3) **SED**: Spectral Error Dressing with degree-only spectrum C_l, the direct comparator for ASED; (4) **IFS-ENS**: ECMWF's operational 50-member ensemble, included as a context reference (note: IFS-ENS uses 00z/12z initialization times vs. GraphCast's 06z/18z).

Metrics. We report the Continuous Ranked Probability Score (CRPS, lower is better), a strictly proper scoring rule for probabilistic forecasts, computed globally and for extra-tropics (30–60° latitude). We also report the Spread-Skill Ratio (SSR), where values closer to 1.0 indicate better calibration. All ensemble methods use $M = 50$ members with 5 random seeds; we report mean $\pm$ std across seeds.

4.2 Anisotropy Diagnostic

Before evaluating ASED, we verify that GraphCast Z500 forecast errors exhibit significant within-degree anisotropy. Computing the calibration anisotropy index on the calibration split yields $A_{\text{cal}} = -0.276 \pm 0.005$, substantially exceeding the isotropy threshold of $|A_{\text{cal}}| < 0.1$. The negative sign confirms quasi-zonal error structure: low-μ (zonal) modes contain $4.26\times$ more power than high-μ (meridional) modes on average.

Figure 2 reveals that anisotropy varies across spatial scales. The synoptic band ($l = 24$–59), corresponding to weather system scales, shows the strongest zonal/meridional ratio of 5.99. Planetary scales ($l = 10$–23) show a ratio of 4.12, while small scales ($l = 147$–359) approach isotropy with a ratio of 1.82. This scale-dependent structure motivates the 4-band decomposition in ASED.

4.3 Main Results

Table 1 presents the main results. ASED achieves a global CRPS of 139.60, representing a 2.92% improvement over SED (143.80). This improvement exceeds the pre-registered 1% significance threshold and is outside the $\pm$std range across seeds, confirming that within-degree anisotropy modeling provides meaningful calibration gains. In the extra-tropics, ASED improves CRPS from 182.90 to 181.54 (0.75% improvement), which is positive but below the 1% threshold. The SSR improves from 1.129 to 1.123 globally, moving closer to the ideal value of 1.0.

Table 1: Probabilistic calibration results on WeatherBench2 Z500 @ 5-day lead time. ASED achieves 2.92% CRPS improvement over SED globally. Best in **bold**, second-best underlined. † indicates reference system with different initialization times.

Method	CRPS Global (↓)	CRPS Extra-tropics (↓)	SSR Global	SSR Extra-tropics
Deterministic	160.93	239.96	–	–
Isotropic GP	143.47±0.74	182.94±0.90	1.127	1.009
SED	143.80±0.52	182.90±0.51	1.129	1.009
ASED (Ours)	**139.60±0.52**	**181.54±0.46**	**1.123**	1.010
IFS-ENS†	117.34	174.75	0.989	**1.010**

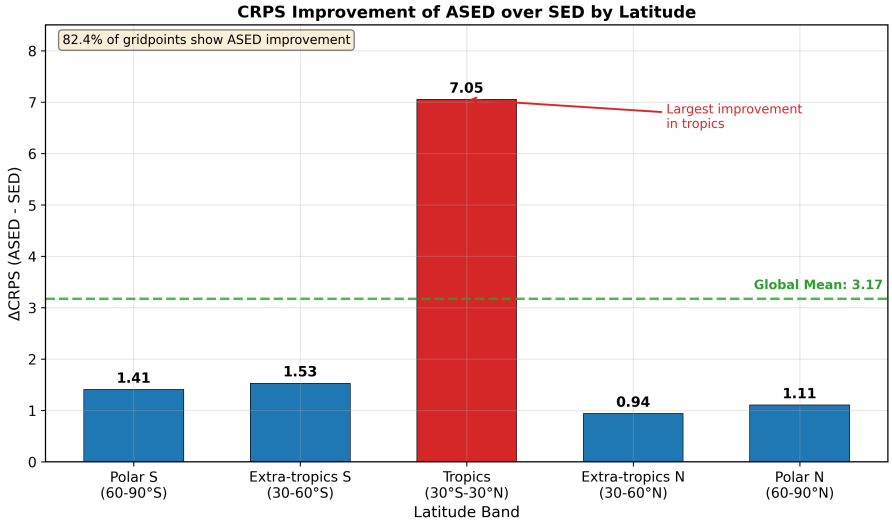

Figure 3: CRPS improvement of ASED over SED by latitude band. ASED improves over SED across all latitude bands (82.4% of gridpoints), with the largest gains in the tropics (ΔCRPS=7.05) and consistent positive gains in extra-tropics (NH: 0.94, SH: 1.53). Green dashed line shows global mean improvement (3.17).

Notably, both SED and Isotropic GP achieve similar CRPS (143.80 vs. 143.47), suggesting that matching the degree spectrum C_l provides comparable benefit to using a fixed isotropic length scale. ASED's improvement over both methods demonstrates the value of modeling within-degree structure. The gap to IFS-ENS (117.34) indicates substantial room for improvement with more sophisticated methods.

4.4 Spatial Analysis

Figure 3 shows the geographic distribution of ASED's improvements. ASED outperforms SED at 82.4% of gridpoints globally, with a mean CRPS improvement of 3.17. Contrary to our initial hypothesis that extra-tropical storm tracks would benefit most, the largest gains occur in the tropics (ΔCRPS = 7.05). The extra-tropics show positive but smaller improvements (Northern Hemisphere: 0.94, Southern Hemisphere: 1.53). This pattern suggests that within-degree anisotropy is particularly pronounced in tropical regions, possibly due to the strong zonal organization of tropical circulation patterns (e.g., trade winds, Hadley cell).

4.5 Limitations

While ASED achieves significant global improvement, the extra-tropics improvement (0.75%) falls below the 1% threshold. This suggests that extra-tropical forecast errors may have additional structure beyond within-degree anisotropy that requires further modeling. Potential directions include latitude-dependent anisotropy profiles, cross-degree correlations, or non-Gaussian error distribu-

tions. Additionally, the gap to IFS-ENS indicates that training-free post-processing, while practical, cannot fully match physics-based ensemble systems.

5 Conclusion

We introduced Anisotropic Spectral Error Dressing (ASED), a training-free method for converting deterministic AI weather forecasts into calibrated probabilistic ensembles. By modeling within-degree anisotropy via μ-binning across 4 degree bands, ASED achieves a 2.92% global CRPS improvement over standard spectral error dressing on WeatherBench2 Z500 forecasts, with 82.4% of gridpoints showing improvement. Our analysis reveals significant quasi-zonal error structure in GraphCast forecasts ($A_{\text{cal}} = -0.276$), with the strongest anisotropy at synoptic scales. While extra-tropics gains are smaller than expected, the method demonstrates that exploiting directional error structure can meaningfully improve probabilistic calibration without model retraining. Future work includes extending ASED to other variables and lead times, modeling latitude-dependent anisotropy, and investigating cross-degree correlations to further close the gap with physics-based ensemble systems.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *Nature*, 619:533–538, 2023.

Z. B. Bouallègue, Tobias Heppelmann, S. Theis, and P. Pinson. Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144:4737–4750, 2016.

C. Bulte, Nina Horat, J. Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*, 2024.

Jian Cao, Jingjie Zhang, Zhuoer Sun, and M. Katzfuss. Locally anisotropic covariance functions on the sphere. In *NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems*, 2022.

Lei Chen, Xiaohui Zhong, Feng jun Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6:1–11, 2023.

Kira Feldmann, M. Scheuerer, and T. Thorarinsdottir. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Monthly Weather Review*, 143:955–971, 2015.

Vignesh Gopakumar, Joel Oskarrson, Ander Gray, L. Zanisi, Stanislas Pamela, Daniel Giles, Matt J. Kusner, and M. Deisenroth. Valid error bars for neural weather models using conformal prediction. *ArXiv*, abs/2406.14483, 2024.

Rémi R. Lam, Alvaro Sanchez-Gonzalez, M. Willson, Peter Wirnsberger, Meire Fortunato, A. Pritzel, Suman V. Ravuri, T. Ewalds, Ferran Alet, Z. Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Jacklynn Stott, O. Vinyals, S. Mohamed, and P. Battaglia. Graphcast: Learning skillful medium-range global weather forecasting. *Science*, 382:1416–1421, 2023.

A. Lang and C. Schwab. Isotropic gaussian random fields on the sphere: Regularity, fast simulation and stochastic partial differential equations. *Annals of Applied Probability*, 25:3047–3094, 2015.

Ankur Mahesh, William Collins, B. Bonev, Noah D. Brenowitz, Y. Cohen, J. Elms, Peter Harrington, K. Kashinath, Thorsten Kurth, Joshua North, Travis OBrien, Michael S. Pritchard, David Pruitt, Mark Risser, Shashank Subramanian, and Jared D. Willard. Huge ensembles part i: Design of ensemble weather forecasts using spherical fourier neural operators. *ArXiv*, abs/2408.03100, 2024.

Jaideep Pathak, Shashank Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, Thorsten Kurth, D. Hall, Zong-Yi Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and Anima Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *ArXiv*, abs/2202.11214, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, T. Ewalds, Andrew El-Kadi, Jacklynn Stott, Shakir Mohamed, Peter W. Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *Nature*, 637:84–92, 2024.

S. Rasp and Sebastian Lerch. Neural networks for post-processing ensemble weather forecasts. *Monthly Weather Review*, 146:3885–3900, 2018.

S. Rasp, P. Dueben, S. Scher, Jonathan A. Weyn, Soukayna Mouatadid, and N. Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12, 2020.

S. Rasp, Stephan Hoyer, Alexander Merose, I. Langmore, P. Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Q. Yang, Rob Carver, Shreya Agrawal, M. Chantry, Z. B. Bouallègue, P. Dueben, C. Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16, 2023.

Roman Schefzik, T. Thorarinsdottir, and T. Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28:616–640, 2013.

Christopher Subich, S. Husain, L. Šeparović, and Jing Yang. Fixing the double penalty in data-driven weather forecasting through a modified spherical harmonic loss function. *ArXiv*, abs/2501.19374, 2025.

Xiaohui Zhong, Lei Chen, Hao Li, Jie Feng, and Bo Lu. Fuxi-ens: A machine learning model for medium-range ensemble weather forecasting. *ArXiv*, abs/2405.05925, 2024.

A Appendix

APPENDIX TEXT