Geosci. Model Dev., 18, 5575–5603, 2025
https://doi.org/10.5194/gmd-18-5575-2025

Geoscientific Model Development
Open Access
EGU

Huge ensembles – Part 1: Design of ensemble weather forecasts using spherical Fourier neural operators

Ankur Mahesh[1,2,★], **William D. Collins**[1,2,★], **Boris Bonev**[3], **Noah Brenowitz**[3], **Yair Cohen**[3], **Joshua Elms**[4], **Peter Harrington**[5], **Karthik Kashinath**[3], **Thorsten Kurth**[3], **Joshua North**[1], **Travis O'Brien**[4], **Michael Pritchard**[3,6], **David Pruitt**[3], **Mark Risser**[1], **Shashank Subramanian**[5], **and Jared Willard**[5]

[1]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory (LBNL), Berkeley, California, USA
[2]Department of Earth and Planetary Science, University of California, Berkeley, USA
[3]NVIDIA Corporation, Santa Clara, California, USA
[4]Department of Earth and Atmospheric Sciences, Indiana University, Bloomington, Indiana, USA
[5]National Energy Research Scientific Computing Center (NERSC), LBNL, Berkeley, California, USA
[6]Department of Earth System Science, University of California, Irvine, USA
★These authors contributed equally to this work.

Correspondence: Ankur Mahesh (ankur.mahesh@berkeley.edu)

Received: 31 July 2024 – Discussion started: 2 October 2024
Revised: 3 April 2025 – Accepted: 6 May 2025 – Published: 4 September 2025

Abstract. Simulating low-likelihood high-impact extreme weather events in a warming world is a significant and challenging task for current ensemble forecasting systems. While these systems presently use up to 100 members, larger ensembles could enrich the sampling of internal variability. They may capture the long tails associated with climate hazards better than traditional ensemble sizes. Due to computational constraints, it is infeasible to generate huge ensembles (comprised of 1000–10 000 members) with traditional, physics-based numerical models. In this two-part paper, we replace traditional numerical simulations with machine learning (ML) to generate hindcasts of huge ensembles. In Part 1, we construct an ensemble weather forecasting system based on spherical Fourier neural operators (SFNOs), and we discuss important design decisions for constructing such an ensemble. The ensemble represents model uncertainty through perturbed-parameter techniques, and it represents initial condition uncertainty through bred vectors, which sample the fastest-growing modes of the forecast. Using the European Centre for Medium-Range Weather Forecasts Integrated Forecasting System (IFS) as a baseline, we develop an evaluation pipeline composed of mean, spectral, and extreme diagnostics. With large-scale, distributed SFNOs with 1.1 billion learned parameters, we achieve calibrated probabilistic forecasts. As the trajectories of the individual members diverge, the ML ensemble mean spectra degrade with lead time, consistent with physical expectations. However, the individual ensemble members' spectra stay constant with lead time. Therefore, these members simulate realistic weather states during the rollout, and the ML ensemble passes a crucial spectral test in the literature. The IFS and ML ensembles have similar extreme forecast indices, and we show that the ML extreme weather forecasts are reliable and discriminating. These diagnostics ensure that the ensemble can reliably simulate the time evolution of the atmosphere, including low-likelihood high-impact extremes. In Part 2, we generate a huge ensemble initialized each day in summer 2023, and we characterize the simulations of extremes.

1 Introduction

Recent low-likelihood, high-impact events (LLHIs) have raised important and unanswered questions about the drivers of these events and their relationship to anthropogenic climate change. For example, Hurricane Harvey in 2017 and the summer 2021 heatwave in the Pacific Northwest (PNW) are two high-impact events with no modern analog. Several threads motivate research on LLHIs. First, the IPCC states

that "the future occurrence of LLHI events linked to climate extremes is generally associated with *low confidence*" (Seneviratne et al., 2021, pp. 1536). Second, the occurrence of recent LLHIs, such as the summer 2021 PNW heatwave, reveals that our abilities to characterize, let alone anticipate, such events are currently incomplete (Bercos-Hickey et al., 2022; Zhang et al., 2024; Liu et al., 2024).

LLHIs challenge the standard climate models that might be used to answer such questions. Computational costs make it infeasible to run the large ensembles of simulations that are necessary to make inferences about the statistics of extremely rare weather events. The climate modeling community has successfully constructed large ensembles of up to $\mathcal{O}(10^2)$ members, such as the Community Earth System Model 2 Large Ensemble (CESM2-LE). To examine the rarest of LLHIs, a larger sample size is necessary. For instance, McKinnon and Simpson (2022) note that "for very large events (e.g., exceeding 4.5σ at a weather station), only a small minority of CESM2-LE analogs in skewness/kurtosis space produce similarly extreme events."

These challenges motivate the application of entirely new methodological approaches, such as those based on machine learning (ML). For the first time, it is now possible to generate massive ensembles using ML with orders-of-magnitude less computational cost than traditional numerical simulations (Pathak et al., 2022). Recent work has demonstrated the potential of deterministic ML-based weather forecasting, which has comparable or superior root mean squared error (RMSE) to the Integrated Forecasting System (IFS) at 0.25° horizontal resolution (Bi et al., 2023; Lam et al., 2023; Willard et al., 2024; ECMWF, 2024). Olivetti and Messori (2024) show that these deterministic data-driven models also offer promising forecast skill on extremes, and Pasche et al. (2024) validate them on case studies of extreme events. As our ML architecture, we use spherical Fourier neural operators (SFNOs) (Bonev et al., 2023). SFNO has been proven to be efficient and powerful in modeling a wide range of chaotic dynamical systems, including turbulent flows and atmospheric dynamics, while remaining numerically stable over long autoregressive rollouts. Given these promising deterministic results, we use ML to create ensemble forecasts, which provide probabilistic weather predictions. A high-level design decision is whether to create the ensemble after training the ML model or during the training itself. We use the former approach: we train ML models to minimize the deterministic mean squared error (MSE) at each time step. After training, we create a calibrated ensemble by representing initial condition and model uncertainty. Conversely, NeuralGCM (Kochkov et al., 2023), FuXi-ENS (Zhong et al., 2024), SEEDS (Li et al., 2024), and GenCast (Price et al., 2023) employ probabilistic training objectives instead of deterministic RMSE.

In this two-part paper, we present a first-of-its-kind huge ensemble of weather extremes using an ML-based emulator of global numerical reanalyses. In Part 1, we introduce the ML architecture and the ensemble design (Sect. 2). In Table 1, we list the major design decisions of the ensemble, and we include pointers to the relevant sections in the paper for understanding the decision-making criteria. We benchmark the ML performance against an operational weather forecast, the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System ensemble (IFS ENS). We assess whether our ML ensemble is fit for purpose using a suite of diagnostics that assess the overall probabilistic performance of the ensemble and its spectra. Because of our interest in LLHIs, we also present an extremes diagnostics pipeline that specifically assesses ensemble extreme weather forecasts. In Part 2 (Mahesh et al., 2025a), we analyze a huge ensemble hindcast with 7424 ensemble members.

2 Designing ensembles with SFNO

We adopt the SFNO training scheme presented in Bonev et al. (2023). SFNO is trained on the European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) (Hersbach et al., 2020) at the dataset's 0.25° horizontal resolution. The weights of SFNO are optimized to minimize the latitude-weighted deterministic MSE loss function. When calculating the loss, each variable is weighted by pressure and by the time tendency of the variable in the training dataset, similar to methods presented in Lam et al. (2023) and Watt-Meyer et al. (2023).

To create ensemble weather forecasts with SFNO, we mirror methods used in numerical weather prediction. Figure 1 provides an overview of our ensemble generation method. For weather forecasts, two major sources of uncertainty are initial condition uncertainty and model uncertainty. Initial condition uncertainty stems from the inaccuracies in observing the current meteorological state, while model uncertainty arises from the incompletely known and imperfect numerical representations of physics that govern the atmosphere's time evolution. To represent initial condition uncertainty, we use bred vectors, a method formerly used by the Global Ensemble Forecast System (GEFS) (Toth and Kalnay, 1993, 1997). Bred vectors are designed to sample the fastest-growing directions of the ensemble error patterns. By creating rapidly diverging ensemble trajectories, bred vectors are designed to create an ensemble that fully represents the probability of future weather states. Existing work has shown that simple Gaussian perturbations do not yield sufficiently dispersive ML ensembles (Scher and Messori, 2021; Bülte et al., 2024): the ensemble spread from these perturbations is too small. Bred vectors help address this problem by creating a more dispersive ensemble which better reflects the full distribution of possible future states of the atmosphere. They amplify the fastest-growing modes in the model's intrinsic dynamics. While bred vectors have been used to create ensemble forecasts from traditional dynamical models, assessing how

Table 1. A list of ensemble design decisions used to create the ML ensemble. The pointer to the section in the paper includes a more in-depth explanation of each decision and the criteria for making the choice.

Name	Value	Paper section
Architecture	Spherical Fourier neural operators v0.1.0	Part 1, Sect. 2.1
Training dataset	1979–2015	Part 1, Sect. 2
Validation dataset	2018	Part 1, Sect. 2
Test dataset	2020	Part 1, Sect. 2
Forecast time step	6 h	Part 1, Sect. 2
Horizontal resolution	0.25°	Part 1, Sect. 2
Embedding dimension	620	Part 1, Sect. 2.1
Scale factor	2	Part 1, Sect. 2.1
Autoregressive fine tuning	None	Part 1, Sect. 3.2
Training time	16 h on 256 A100 GPUs per checkpoint	Part 1, Sect. 2.1
Inference time	1 s per 6 h time step on 1 NVIDIA A100 GPU	Part 1, Sect. 2.1
Variable set	73 channels from Bonev et al. (2023) and 2 m dew point temperature. The pressure variables are represented on 13 pressure levels.	Part 1, Sect. 2.1
	0.35* SFNO deterministic RMSE at 48 h	Part 1, Sect. 2.3
Centered bred vectors	Each bred vector is added to and subtracted from the initial condition	Part 1, Sect. 2.3
Hemispheric rescaling for bred vectors	Perturbations are rescaled separately polewards of 20°. A linear interpolation is used for rescaling in the tropics.	Part 1, Sect. 2.3
Initial noise for bred vectors	Adding spherical noise (correlated on 500 km length scales) to z_{500} (500 hPa geopotential)	Part 1, Sect. 2.3
Number of model checkpoints	29	Part 1, Sect. 2.2
Number of perturbations per model checkpoint (benchmark ensemble)	2 (1 bred vector perturbation that is added to and subtracted from the initial condition)	Part 1, Sect. 3
Number of perturbations per model checkpoint (huge ensemble)	256 (128 bred vector perturbations, each added to and subtracted from the initial condition)	Part 2
Total size of huge ensemble	7424 members	Part 2
Lead time to analyze extreme statistics	2, 4, 10 d	Part 1, Sect. 3.3
Derived variables in huge ensemble	Integrated vapor transport, 10 m wind speed, heat index	Part 2

ML models respond to such perturbations is an important research frontier.

To represent model uncertainty, we train multiple SFNO models from scratch. We refer to each trained SFNO instance as a "checkpoint". At the start of training, each checkpoint is initialized with different random weights. During training, SFNO iteratively updates its weights to minimize a loss function: in this case, the loss function is the mean squared error between the model predictions and the ERA5 training data. During each epoch of training, SFNO iterates through the entire training dataset and updates its weights to minimize the loss. We train SFNO for 70 total epochs. By the end of training, the models converge to a different local optimum of learned weights. The resulting ensemble of the different trained SFNO checkpoints represents the uncertainty in the SFNO model weights itself. Each resulting checkpoint represents an equivalently plausible set of weights that can model the time evolution of the atmosphere from an ini-

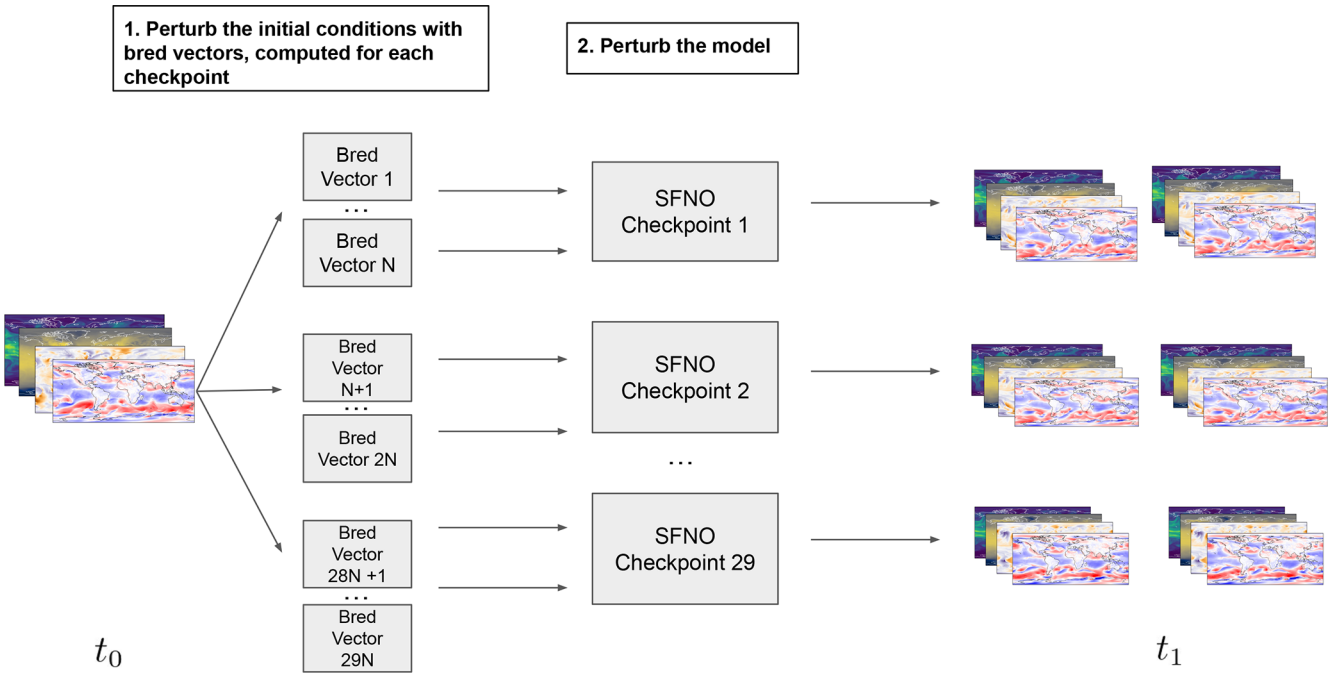


Figure 1. Overview of ensemble architecture. The ensemble is constructed using two methods: initial condition perturbations and model perturbations. The initial condition perturbations are generated using bred vectors to sample the fastest-growing errors in the initial condition. Model perturbations consist of 29 instances of the SFNO model trained independently from scratch. Bred vectors are generated separately for each SFNO checkpoint. Each bred vector creates two initial condition perturbations: one with the bred vector added to the initial condition and one with the bred vector subtracted from the initial condition. For the small ensemble, we use $N = 1$ bred vectors per checkpoint. For the huge ensemble in Part 2, there are $N = 128$ bred vectors per checkpoint.

tial state. With multiple checkpoints, we create an ensemble with a spread of forecasts. Weyn et al. (2021) use multiple checkpoints to create an ensemble of forecasting models for medium-range and subseasonal prediction. They reduce computational costs by saving multiple checkpoints from each training run and training the last few epochs independently for each model. This approach requires several additional design decisions: how should the learning rate for the optimization during these last retrained epochs be adjusted? How many extra epochs should each checkpoint train for? At what point during training should the checkpoints diverge? To minimize the ensemble's dependence on these hyperparameters, we opt to retrain each checkpoint completely from scratch.

We create an ensemble called SFNO-BVMC: spherical Fourier neural operators with bred vectors and multiple checkpoints. In Table 1, we present a list of hyperparameters and their associated criteria that we use to guide our choice of ensemble design. We use a train–validation–test set paradigm. SFNO is trained on the years 1979–2016. We use the year 2018 as a validation year, on which we tune multiple aspects of the ensemble, such as the amplitude of the bred vectors and the number of SFNO checkpoints. Because these ensemble parameters are tuned using the year 2018, we cannot use 2018 for unbiased evaluation of the final ensemble. For our overall diagnostics, the year 2020 is used as an out-of-sample, held-out test set. To evaluate the skill for extreme weather, we use boreal summer 2023 (June, July, August) because it was the hottest summer in recorded history at the time (Esper et al., 2024). In Part 2, we present a deep dive on a huge ensemble of forecasts from this time period. No SFNO training or ensemble design decisions were made using the year 2020 or 2023. This setup with different training, validation, and test sets is crucial to avoid data leakage.

2.1 Selecting an emulator

SFNO is an ML architecture built on neural operators (Li et al., 2020), which are designed to learn mappings between function spaces. They can be used for different discretizations and grids, and they have broad applicability to various partial differential equation (PDE) problems. SFNO is a special instance of a neural operator, which uses the spherical harmonic transform to represent operators acting on functions defined on the sphere. The spherical formulation leads to a strong inductive bias, respecting the geometry and symmetry of the sphere. This reduces error buildup during autoregressive rollouts. We use the open-source version of SFNO v0.1.0 released in the modulus-makani Python repository (Bonev et al., 2024).

We provide a brief overview of the SFNO architecture; for a more detailed explanation, refer to Bonev et al. (2023). The SFNO architecture consists of three main components: the encoder, the SFNO blocks, and the decoder.

1. Encoder: the encoder employs multi-layer perceptrons (MLPs) at each grid cell to map the input fields into a higher-dimensional latent space. MLPs are fully connected neural networks that apply nonlinear transformations to their inputs.

2. SFNO blocks: the processor incorporates eight SFNO blocks, operating in the latent space, each performing two main operations as follows.

    a. A spherical convolution with a learned filter encoded in the spectral domain. The signal is transformed into the spectral domain and back via a spherical harmonic transform (SHT) and its inverse. In the spectral domain, the convolution operation becomes a pointwise multiplication.

    b. An MLP applies nonlinear transformations to the latent features.

    The output of each SFNO block serves as the input to the subsequent block. The first block downsamples the input resolution by a specified "scale factor", while the last block upsamples back to the original resolution.

3. Decoder: the decoder maps the latent space back into physical space using MLPs.

SFNO encodes an operator that maps functions defined on the sphere to other functions on the sphere. This learned map is parameterized by the weights of the MLPs, spectral filters, encoder, and decoder. These weights of SFNO are optimized during training.

The input to SFNO consists of 74 channels comprising the meteorological state at a given time (Table 2). The model then predicts those same 74 channels 6 h in the future. In addition to the prognostic channels, we add three extra input channels: the cosine of the solar zenith angle, orography, and land–sea mask.

The existing implementation of SFNO from Bonev et al. (2023) makes forecasts for 73 total prognostic channels. In this study, we add ERA5 2 m dew point temperature as another variable. Together, 2 m dew point temperature and 2 m air temperature provide an estimate of heat and humidity at the surface. Since we have trained SFNOs to predict both these variables, we can simulate LLHI heat–humidity events. It is vital to assess the combination of both heat and humidity to characterize heat stress and LLHIs in a warming world (Vargas Zeppetello et al., 2022). While some ML models have 1000 hPa specific humidity, Pasche et al. (2024) note that this approximation has limitations in predicting the surface heat stress and heat index. Therefore, we add 2 m dew point to more directly characterize moisture near the surface. In future work, the addition of 2 m dew point could enable estimating convective available potential energy in the forecasts from SFNO. By quantifying the buildup of convective instability, this variable is useful for studying convective storms and thunderstorms.

The SFNO architecture includes scalable model parallel implementations, in which the model is split across multiple GPUs during training (Bonev et al., 2023; Kurth et al., 2023). We train large SFNOs and assess the effect of the SFNO size on ensemble dispersion. SFNO contains a number of hyperparameters that determine the total size of the model and its ensemble performance. Two such hyperparameters are the scale factor and the embedding dimension. The scale factor specifies how much the input field is spectrally downsampled when creating the latent representation. With more aggressive downsampling, SFNO internally represents the input atmospheric state with reduced resolution. We speculate that this may reduce the effective resolution of the predictions (Brenowitz et al., 2024). With a reduced effective resolution, small-scale perturbations would not grow and propagate upscale. Instead, they would be blurred out, and they would not result in increased spread among ensemble members. The embedding dimension determines the size of the learned representation of the input fields (Pathak et al., 2022). A larger embedding dimension increases the number of learnable parameters in the SFNO, thereby requiring more GPU memory.

We compare three combinations of these hyperparameters: a *small* SFNO, a *medium* SFNO, and a *large* SFNO. The small SFNO has a scale factor of 6 and embedding dimension of 220, the medium-sized model has a scale factor of 4 and embedding dimension of 384, and the large model has a scale factor of 2 and embedding dimension of 620. The small, medium, and large SFNOs have 48 million learned weights, 218 million learned weights, and 1.1 billion learned weights, respectively. Based on the number of weights, the large SFNOs are among the largest ML-based weather forecasting models currently available.

To select an SFNO architecture, we assess how these hyperparameters affect lagged ensemble spread–error ratio and spectral degradation. A lagged ensemble is created by using nine adjacent time steps as initial conditions (Brankovic et al., 1990). Brenowitz et al. (2024) analyze the spread–error ratio of lagged ensembles to assess the intrinsic dispersion of deterministic ML models. Ordinarily, benchmarking ensemble performance would require generating and tuning a full set of ensemble parameters (e.g., amplitude of perturbations, number of checkpoints, form of perturbations) separately for each architecture. This process is time-consuming, memory-intensive, and computationally demanding. Lagged ensembles readily enable comparison of different deterministic architectures with minimal tuning parameters. In Fig. 2a, the lagged ensemble spread–error ratio for 850 hPa temperature is closest to 1 for the large model, indicating that this model

Table 2. A list of the pressure, surface, and static variables used by SFNO. All variables are used as input to the model, whereas only prognostic variables are predicted.

Type	Variable	Pressure levels (hPa)	Prognostic
Atmospheric variables	temperature	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50	√
	specific humidity	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50	√
	geopotential	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50	√
	zonal wind	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50	√
	meridional wind	1000, 925, 850, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50	√
Surface variables	2 m air temperature	surface	√
	2 m dew point temperature	surface	√
	total column water vapor	surface	√
	surface pressure	surface	√
	sea level pressure	surface	√
	10 m zonal wind	surface	√
	10 m meridional wind	surface	√
	100 m zonal wind	surface	√
	100 m meridional wind	surface	√
Misc.	cosine zenith angle	–	x
	orography	–	x
	land–sea mask	–	x

is best-suited for ensemble forecasting. The spread–error ratio systematically improves for the larger models. Brenowitz et al. (2024) find complementary results; they show that smaller scale factors improve the spread–error ratio. Here, we consider the combined effect of changing both the scale factor and embedding dimension.

We assess the extent to which the small, medium, and large SFNOs fully resolve the spectrum of the ERA5 training data. A known problem with deterministic ML models is that the small wavelengths are blurry (Kochkov et al., 2023). We attempt to suppress this blurring as much as possible by using a small scale factor and a large embedding dimension. In addition, we intentionally avoid using autoregressive training (Lam et al., 2023; Pathak et al., 2022; Keisler, 2022). In this method (sometimes called "multistep fine tuning" or "multistep loss"), the ML model weights are optimized over multiple time steps, not just a one-step prediction. The goal of this method is to improve the forecast performance by training the ML model to perform well when autoregressively rolled out with its own predictions. Brenowitz et al. (2024) and Lang et al. (2024) hypothesize that autoregressive training could contribute to spectral degradation. This method may effectively increase the time step of the model, making it more similar to an ensemble mean (Lang et al., 2024). Many deterministic models' initial one-step forecasts are blurry, and with this method, their forecasts get increasingly blurry with lead time. Because we do not use autoregressive fine tuning, we hypothesize that SFNO has spectra that stay constant with lead time (see Sect. 3.2 for more discussion).

With this design decision, we also reduce the computational requirements of training SFNO. Autoregressive fine tuning requires significant GPU memory and computation time because it calculates gradients across multiple model steps. With these computational savings, we train large SFNOs with a small scale factor and a large embedding dimension. These design choices allow our configuration of SFNO to hold as much high-resolution information in its internal representation as possible.

Figure 2b shows that the larger models (with lower scale factors and larger embed dimension) have less spectral degradation and are better able to preserve the spectra of ERA5. Based on Fig. 2a and b, we select the large version of SFNO as our final set of hyperparameters. This version of SFNO trains in 16 h on 256 80 GB NVIDIA A100 GPUs. It leverages data parallelism, in which the batch size of 64 is split up across different GPUs, and spatial model parallelism, in which the input field is divided into four latitude bands. Each SFNO checkpoint trains for 70 epochs using a pressure-weighted mean squared error loss function (Lam et al., 2023).

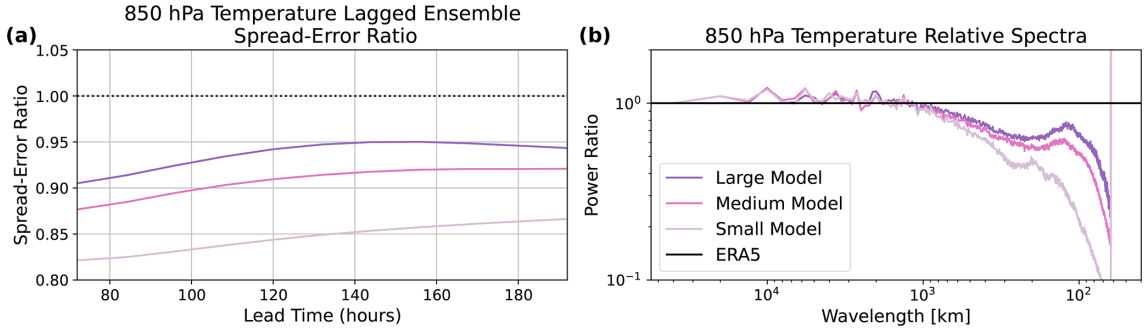


Figure 2. Comparing different versions of SFNO. **(a)** The 850 hPa temperature spread–error ratios are compared for lagged ensembles. A lagged ensemble is created by using nine adjacent time steps as initial conditions, and the spread–error is shown for each SFNO configurations. **(b)** Relative power spectra at a lead time of 360 h (colored lines) for 850 hPa temperatures for a large SFNO (with a scale factor of 2 and an embed dimension of 620), a medium-sized SFNO (scale factor 4 and embed dimension 384), and a small SFNO (scale factor 6 and embed dimension 220). Spectra are computed relative to the ERA5 spectrum (horizontal black line).

We note that there are many possible combinations of the scale factor, embedding dimension, and other training hyperparameters. We do not conduct comprehensive hyperparameter tuning via a grid search. Such an experiment would be very computationally expensive due to the large number of combinations. Instead, we optimize the scale factor and embedding dimension because of their direct relevance to spectral degradation. Rather than hyperparameter tuning, we choose to expend our compute budget on training as many checkpoints as possible to try to span the space of all possible SFNO checkpoints. With many SFNO checkpoints, we hope to increase our coverage of extreme weather events with a thorough representation of model uncertainty.

2.2 Selecting a number of checkpoints for the ensemble

We determine that 29 checkpoints adequately sample the ensemble spread. We experiment with using different numbers of checkpoints in the size of the ensemble, from 4 checkpoints to 34 checkpoints, at intervals of 5 checkpoints. For each ensemble size, we conduct 100 bootstrap samples with replacement from the 34 checkpoints. Figure 3 shows the resulting ensemble spread obtained from these bootstrap samples. The ensemble spread is calculated as the global mean ensemble variance at each grid cell; it is calculated for a 120 h lead time and averaged over forecasts initialized at 52 initialization dates (one initialization per week of 2018). We choose 120 h because this timescale allows synoptic-scale errors to grow, and given its importance for weather forecasting, we hope to represent model uncertainty for this time period as accurately as possible. Figure 3 shows that the ensemble spread becomes an asymptote at approximately 29 checkpoints. We conclude that 29 checkpoints adequately sample the underlying population of all possible SFNO checkpoints with our selection of hyperparameters. In our ensemble results for the remainder of this paper, we use 29 checkpoints. We open-source all 34 model checkpoints (each with 1.1 billion learned weights) as a resource to the community to explore the benefit of multiple SFNOs on forecasting atmospheric phenomena.

2.3 Bred vectors with SFNO

Bred vectors are a computationally efficient way to sample the fastest-growing modes of the atmosphere (Toth and Kalnay, 1993). In Fig. 4, we generate bred vectors using the following methodology.

1. Generate spherical random noise correlated on 500 km length scales. Add this noise as a perturbation to 500 hPa geopotential at time t_{-2}.

2. Generate a perturbed forecast (with the perturbed input) and a control forecast (with the unperturbed input).

3. Subtract the control forecast from the perturbed forecast. Use this difference as the perturbation. (Unlike the initial noise in step 1, this perturbation is applied to all variables and pressure levels, not just z_{500}.)

4. Rescale the perturbation in each hemisphere to the target amplitude of the perturbation.

5. Repeat steps (2)–(4) for t_{-1} and t_0.

The resulting perturbation can be added to or subtracted from t_0. Using this perturbed initial condition, SFNO generates a 360 h forecast, which serves a perturbed member in the ensemble.

The amplitude of the bred vectors is determined by the deterministic RMSE of SFNO at 48 h, multiplied by a scaling factor of 0.35. This factor is a tuning parameter. Since this parameter is less than 1, it reduces the perturbation amplitude. At early lead times, the deterministic and ensemble mean RMSEs of an ensemble forecast are similar. To satisfy criteria for statistical exchangeability, the ensemble spread

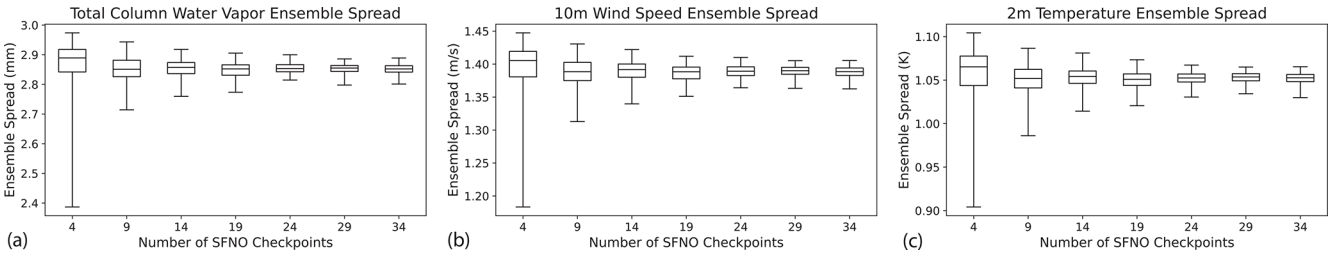


Figure 3. Ensemble spread from different numbers of checkpoints. Ensemble spread is calculated as the square root of time mean, global mean variance (Fortin et al., 2014). A correction factor of $N-1$ is applied to account for different ensemble sizes in the unbiased estimator of variance. At a lead time of 5 d, ensemble spread is averaged over forecasts from 52 initial conditions in the validation set (one per week starting 1 February 2018). Ensemble spread is shown for total column water vapor **(a)**, 10 m wind speed **(b)**, and 2 m temperature **(c)**. For each number of SFNO checkpoints, 200 estimates of ensemble spread are obtained by taking 100 bootstrap random samples of the SFNO checkpoints. The boxes and whiskers visualize the distribution of these 200 trials: the middle of the box is the median, the ends of the box are the first and third quartile of the data, and the ends of the box correspond to the minimum and maximum.

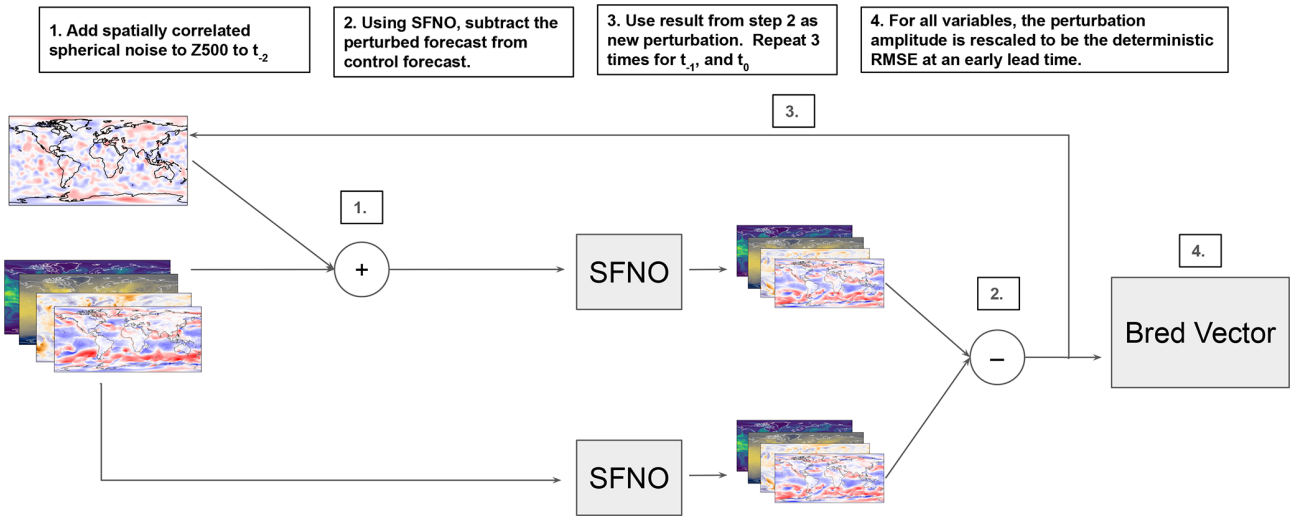


Figure 4. Diagram of generating bred vectors. This diagram details the process of generating bred vectors used for developing initial condition perturbations at t_0. First, using the input three time steps before t_0 (denoted t_{-2}), random noise is added to 500 hPa geopotential (z_{500}). This noise respects spherical geometry and has a spatial correlation length scale of 500 km. With t_{-2} as the initial condition, the perturbed forecast is subtracted from the control forecast. This difference is rescaled and used as a new perturbation, which is added to t_{-1}. This process is repeated for t_0. For each variable during every step of the breeding process, the amplitude of the perturbation is scaled to be 0.35 times the deterministic RMSE of SFNO at 48 h.

should match its ensemble mean RMSE. We use the deterministic RMSE (with a tuning parameter) as a proxy for the desired spread level at early lead times. This approach provides a clear guide for the amplitude of each variable at each pressure level. Manually tuning these amplitudes across variables and levels would be challenging, since there are 74 different input variables. Figure B1 shows the actual amplitude for each of the 74 variables.

We adopt two design choices from Toth and Kalnay (1997) and Toth and Kalnay (1993): centered perturbations and hemispheric-dependent amplitudes. For each learned bred vector, we both add it to and subtract it from the initial condition; this creates two separate perturbations. Centered perturbations improved the performance of the ensemble mean RMSE on the 2018 validation set. Additionally, we rescale the amplitudes separately for the Northern Hemisphere extratropics and the Southern Hemisphere extratropics. To prevent jump discontinuities in the perturbation amplitudes near 20° N and 20° S, a linearly interpolated rescaling factor is used in the tropics. Hemispheric rescaling prevents one hemisphere from dominating the perturbation amplitude. All perturbations are clipped to ensure that total column water vapor and specific humidity cannot be negative. See Appendix E for a note about our implementation of bred vectors.

In step 2 of Fig. 4, we add correlated spherical noise to 500 hPa geopotential (z_{500}). The noise has a correlation length scale of 500 km, and it has the same structure as noise of the stochastic perturbed parameterized tendency scheme

used at ECMWF (Leutbecher and Palmer, 2008). We only add the initial noise to z_{500} to avoid perturbing different fields in opposing and possibly contradictory directions. For instance, positively perturbing total column water vapor but negatively perturbing specific humidity on the lower pressure levels would likely be unphysical. z_{500} is a natural choice of initial field to perturb because it is the steering flow in the extratropics. Since it is a smooth field on an isobaric surface, correlated spherical noise is an appropriately structured additive perturbation. On the other hand, correlated spherical noise would not serve well as an additive perturbation to surface fields, which have sharp discontinuities due to orography and land–sea contrasts. We design the bred vectors with the goal of keeping the perturbed input as close to the training dataset as possible. We minimize the extent of directly prescribed perturbations, and the majority of the perturbation structure is generated from the breeding process with SFNO itself. To start the breeding cycle, the initial perturbation is applied to z_{500}, but for all subsequent cycles, all 74 input variables are perturbed. In this manner, we develop a mutually consistent way of perturbing all input channels.

We test our bred vectors by evaluating spread–error performance on the validation year: 2018. Figure 5 visualizes sample bred vectors for various input fields and channels. These perturbations contain some desirable qualities. First, they contain a land–sea contrast for surface fields such as 10 m wind speed and 2 m temperature. In this example, the 2 m temperature perturbation has an amplitude of 0.56 K over land and 0.27 K over the ocean, and the 10 m wind speed perturbation has an amplitude of $0.45\,\mathrm{m\,s^{-1}}$ over land and $0.66\,\mathrm{m\,s^{-1}}$ over the ocean. The specific humidity perturbations are stronger in the tropics than at the poles, in line with the Equator-to-pole moisture gradient. These physical qualities of bred vectors are a benefit of using bred vectors over simple spherical noise, as in GraphCast-Perturbed (Price et al., 2023) or Perlin noise (Bi et al., 2023).

We initially presented bred vectors and multiple checkpoints in Collins et al. (2024). Concurrently, Baño-Medina et al. (2024) released a preprint using bred vectors and multiple trained models. The results in Baño-Medina et al. (2024) serve as excellent independent validation of bred vectors and multiple checkpoints. They validate their method from 10 January to 28 February (with 50 forecast initial dates), and they show promising results, particularly at certain latitudes and land regions. We comprehensively show that SFNO-BVMC is competitive with IFS on global mean quantities using forecasts from a full year (732 forecast initial dates for 2020 and 92 for summer 2023). We further validate our ensemble with a unique pipeline for extreme diagnostics and spectral diagnostics of each ensemble member and the ensemble mean. While their method uses adaptive Fourier neural operators (AFNOs) (Pathak et al., 2022), we use SFNO, a successor to AFNO that is more stable and has better skill. We train all 29 SFNOs from scratch, whereas they sample multiple models from three training runs. To compare methodologies with Fig. 2 in Baño-Medina et al. (2024), we present a diagram of how we generated bred vectors in SFNO-BVMC. The boxed quantities in Fig. 4 represent the unique methodological details of our approach. We add spherical initial noise to z_{500} (compared to Gaussian noise), start the breeding cycle three time steps before the initialization date (compared to 1 January 2018), and use the deterministic RMSE as the bred vector amplitude. In Part 2, we assess the forecasts from bred vectors and multiple checkpoints at scale, with a significantly larger ensemble than in Baño-Medina et al. (2024).

2.4 Contributions of bred vectors and multiple checkpoints to the ensemble calibration

In SFNO-BVMC, the bred vectors and multiple trained model checkpoints both contribute to ensemble spread and calibration. Bred vectors are a flow-dependent initial condition perturbation: they are calculated independently for each checkpoint, and they use the preceding two time steps to generate the perturbation based on the current flow in the atmosphere. At longer lead times, when there is less dependence on the initial conditions, multi-checkpointing causes the spread–error ratio to approach 1; this is consistent with our expectations from the ensemble in Weyn et al. (2021). In Fig. 6, we show the spread–error ratios from three different ensembles: Fig. 6a has only 29 checkpoints and no bred vectors, Fig. 6b has 1 checkpoint and 29 bred vectors (each added to and subtracted the initial condition), and Fig. 6c has 29 checkpoints and 1 bred vector (added to and subtracted from the initial condition). Figure 6a has 29 ensemble members, while Fig. 6b and c have 58-member ensembles. 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 timescales (3–5 d) when representing model uncertainty also becomes important for obtaining good calibration.

3 Ensemble diagnostics

Ultimately, with SFNO-BVMC, we hope to study LLHIs. This requires a calibrated ensemble with reliable probabilistic forecasts. SFNO-BVMC is a novel way to create ensemble forecasts from deterministic ML models. Therefore, in the following section, we present a diagnostics pipeline to evaluate the SFNO-BVMC ensemble and compare it to the IFS ensemble. We first evaluate SFNO-BVMC using diagnostics that evaluate overall performance. Next, we assess SFNO-BVMC's control, perturbed, and ensemble mean spectra. Finally, we present diagnostics specifically focused on extreme weather forecasts. We open-source the code for these diagnostics (see the "Code and data availability" sec-

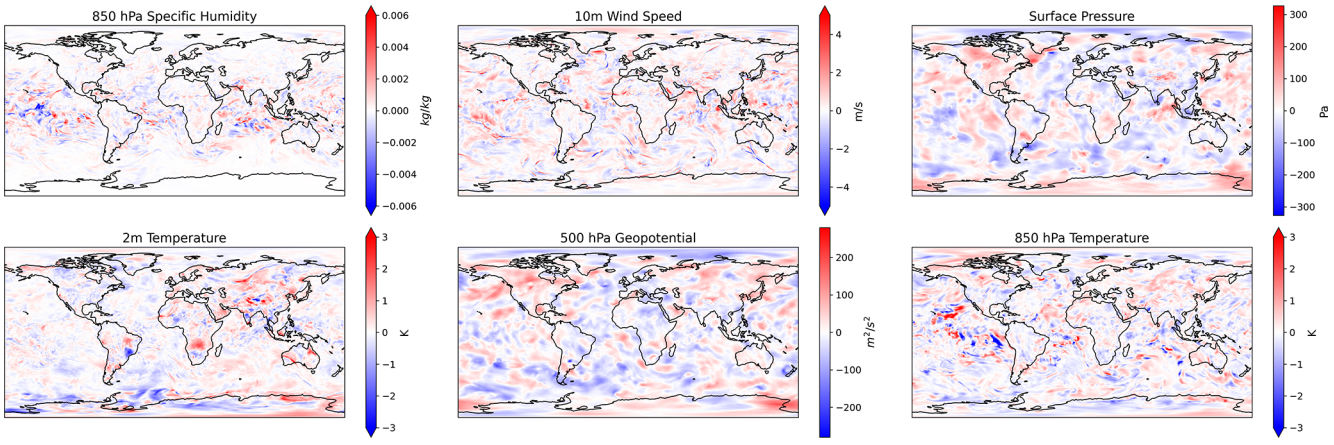


Figure 5. Sample visualizations of the learned bred vectors. For a sample initial time (18 June 2020, 00:00 UTC), the bred vectors are visualized for six different input fields: 850 hPa specific humidity, 10 m wind speed, surface pressure, 2 m temperature, 500 hPa geopotential, and 850 hPa temperature.

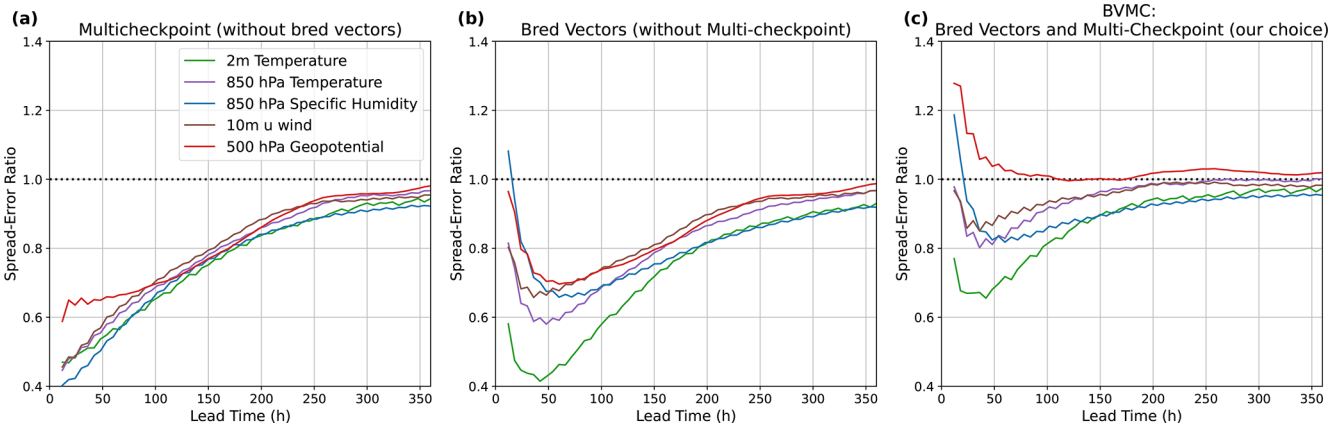


Figure 6. Contributions of bred vectors and multiple checkpoints to spread–error relations. Panel **(a)** shows the spread–error relation obtained from an ensemble only composed of multiple checkpoints. This ensemble has 29 members, one for each checkpoint. Panel **(b)** shows the same for an ensemble of 58 members, using only bred vectors for initial condition perturbations. Panel **(c)** shows the spread–error relation for an ensemble composed of 58 members, with one bred vector added and subtracted from the initial condition for each model checkpoint. Spread–error ratios are averaged across 52 initial conditions, one per week starting 1 February 2018. Successful ensemble forecasts have a spread–error ratio of 1.

tion), and we hope that it can be used to guide future ML model development. For a fair comparison for all diagnostics, we validate IFS against ECMWF's operational analysis and SFNO-BVMC against ERA5. IFS is initialized with this operational analysis, not the ERA5 reanalysis, so it has a different verification dataset. All diagnostics show SFNO-BVMC scores with 58 members and IFS ENS scores with 50 members. SFNO-BVMC has 58 members: 29 checkpoints and 1 bred vector per checkpoint (added to and subtracted from the initial condition). Because of the use of 29 checkpoints and centered bred vector perturbations, SFNO-BVMC cannot be evaluated with an ensemble size smaller than 58 members. While there are versions of the metrics that are corrected for ensemble size, the difference in the metrics due to different ensemble size would be sufficiently small that the diagnostics still allow for comparison between the 50-member IFS and 58-member SFNO-BVMC.

3.1 Mean diagnostics

We validate the overall quality of the ensemble on three diagnostics: continuous ranked probability score (CRPS), spread–error ratio, and ensemble mean RMSE. First, CRPS evaluates a probabilistic forecast of a ground-truth value. It is a summary score of the performance of the ensemble fore-

cast. The formula for CPRS at a given grid cell is

$$\begin{aligned}\mathrm{CRPS}(F, y) &= \int_{-\infty}^{\infty} (F(z) - 1\{y \le z\})^2 \mathrm{d}z \\ &= E_F|X - y| - \frac{1}{2} E_F|X - X'|, \end{aligned} \tag{1}$$

where X and X' are random variables drawn from the cumulative distribution function (CDF) of the ensemble forecast F. Here, y is the verification value (ERA5 for SFNO-BVMC and operational analysis for IFS ENS).

Figure 7 compares the global mean CRPS of SFNO-BVMC to that of IFS ENS on five different variables. On 850 hPa temperature, 2 m temperature, 850 hPa specific humidity, and 500 hPa geopotential, SFNO-BVMC lags approximately 12–18 h behind IFS ENS. SFNO-BVMC does match IFS ENS on the 10 m zonal (u component) wind.

Second, an essential requirement for an ensemble weather forecast is that the ensemble spread must match its skill (Fortin et al., 2014); the spread–error ratio should be 1. This result is derived statistically based on the idea of exchangeability between ensemble members: each ensemble member should be statistically indistinguishable from each other and from the forecasts (Fortin et al., 2014; Palmer et al., 2006). The spread is the square root of the global mean ensemble variance. Similarly, the error is the square root of the global mean ensemble MSE. See Appendix C for a detailed description of calculating the spread and error across multiple forecasts initialized on different initial times. Figure 8 demonstrates that SFNO-BVMC obtains spread–error ratios that approach 1, and it has comparable performance to IFS ENS. At early lead times, SFNO-BVMC is underdispersive for all variables except z_{500}, but the spread skill ratio approaches 1 for longer lead times.

Finally, we evaluate the ensemble mean RMSE of SFNO-BVMC and IFS ENS (Fig. 9). Their scores are comparable, with SFNO-BVMC lagging close behind the IFS ensemble mean, and both models have an ensemble mean RMSE that converges to climatology at 360 h (14 d).

Through large SFNOs with a high-resolution expressive internal state, bred vectors, and multi-checkpointing, this ensemble has significantly improved calibration compared to previous work using lagged ensembles (Brenowitz et al., 2024). It serves as a benchmark for the calibration potential for deterministic ML models, and it can be compared to recent models which optimize for an ensemble objective. While IFS ENS has been an established weather forecasting model for decades, SFNO is still a new architecture. Improving the skill of the SFNO architecture itself is an important area of future research. However, in this paper, our main goal is not primarily to create the most skillful weather forecasting model; rather, we hope to explore huge ensembles and low-likelihood events at the tail of the ensemble forecast distribution. SFNO-BVMC is less computationally expensive than IFS, so it uniquely enables the creation of huge ensembles. These allow for unprecedented sampling of internal variability and an analysis of extreme statistics, as presented in Part 2 of this paper.

3.2 Spectral diagnostics

A common issue with deterministic machine learning weather models is that their forecasts tend to be "blurry" (Kochkov et al., 2023). As a metric to measure and quantify this blurriness, existing work compares the spectra of the ML predictions to the spectra of ERA5. The spectral analyses show that ML models have reduced power at small wavelengths compared to ERA5. Deterministic ML models are often trained using the MSE loss function, which strongly penalizes sharp forecasts in the wrong place. This is referred to as the double penalty problem (Mittermaier, 2014), in which an ensemble is penalized once for predicting a storm in the wrong place and another time for missing the correct location of the storm. To avoid the double penalty from the mean squared error, ML models may learn to predict smooth, blurred solutions that appear closer to an ensemble mean (Agrawal et al., 2023; Brenowitz et al., 2024), rather than an individual ensemble member.

Regarding spectral performance, there are two desirable characteristics. These characteristics distinguish an individual ensemble member from an ensemble mean.

1. During the rollout, it is preferable for the spectra of each ensemble member to stay constant with lead time. With this characteristic, each ensemble member maintains a realistic representation of the atmospheric state during the rollout.

2. During the rollout, it is preferable for the spectra of the ensemble mean to realistically degrade with lead time (Bonavita, 2023). As the ensemble members spread more and their trajectories diverge, the ensemble mean should become blurrier. In particular, on synoptic timescales (around 3–5 d), when error growth becomes nonlinear, the IFS ensemble mean displays a sharp decline in power around 1000 km wavelengths (Bonavita, 2023).

On the first characteristic, SFNO-BVMC spectra remain constant from 24 to 360 h (Figs. 10 and 11). This contrasts with GraphCast and AIFS; those deterministic ML models do increasingly blur with lead time (Kochkov et al., 2023; Lang et al., 2024). We hypothesize that SFNO-BVMC spectra remain constant because of our intentional choice not to use autoregressive training.

Through this test, we verify that the individual members' predictions do not collapse into the ensemble mean. This is a crucial test of the physical fidelity of SFNO-BVMC. Because each SFNO-BVMC ensemble member's spectrum is constant through the rollout, the ensemble members maintain their ability to resolve extreme weather. If their spectra

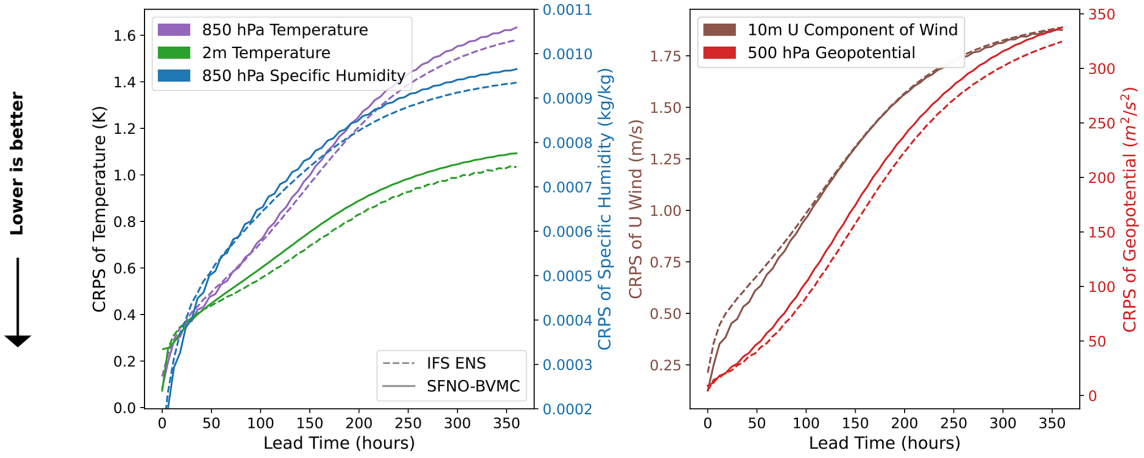


Figure 7. CRPS of SFNO-BVMC and IFS ENS. SFNO-BVMC is a 58-member ensemble that uses 29 SFNO checkpoints trained from scratch and two initial condition perturbations per checkpoint. The two initial condition perturbations come from a single bred vector that is added to and subtracted from the initial condition. Scores are calculated over 732 initial conditions (two per day at 00:00 and 12:00 UTC) for 2020, which is the test set year. SFNO-BVMC is validated against ERA5, and IFS ENS is validated against ECMWF's operational analysis. IFS ENS scores are taken from WeatherBench 2 (Rasp et al., 2024).

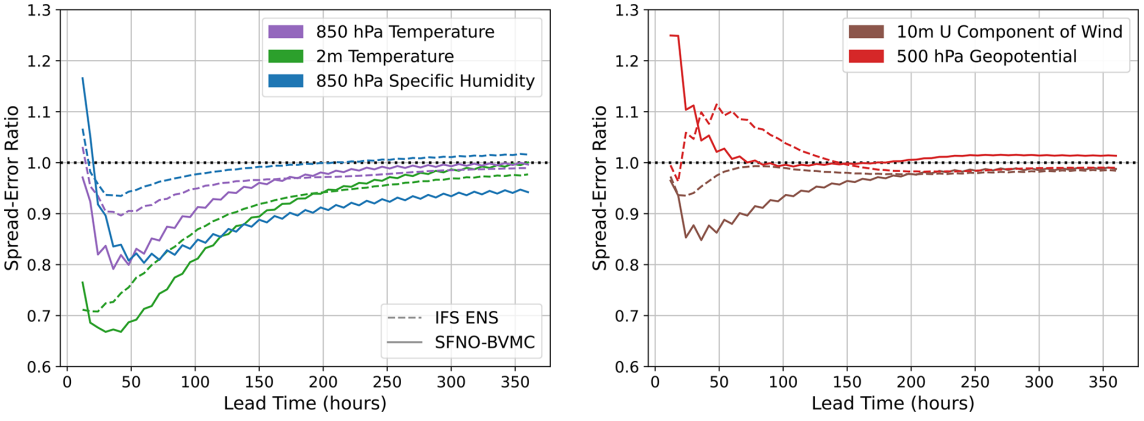


Figure 8. Spread–error ratio of SFNO-BVMC and IFS ENS. SFNO-BVMC is the same 58-member ensemble described in Fig. 7. Spread–error ratios are calculated over 732 initial conditions (two per day at 00:00 and 12:00 UTC) for 2020. SFNO-BVMC is validated against ERA5, and IFS ENS is validated against ECMWF's operational analysis. IFS ENS scores are taken from WeatherBench 2 (Rasp et al., 2024).

degraded with lead time, then the forecasts may become too blurry to predict highly localized extreme events. At a lead time of 360 h, the perturbed members maintain similar spectra as the control member (Fig. 11), and at the initial time, they have similar spectral characteristics as the unperturbed ERA5 initial condition (Fig. D5).

An important caveat is that even though the spectra are constant during the rollout, they are still somewhat degraded compared to ERA5 (Fig. 2b). This degradation occurs in the first 24 h of the rollout. We have not solved the problem of blurry forecasts entirely. We have minimized it as much as possible by using a large embedding dimension and a small scale factor, which increase the resolution of the latent representation of the input, and by intentionally avoiding multistep fine tuning. However, our deterministic training setup still results in blurring with the use of the MSE loss function and large 6 h time steps, and alleviating this problem is an important avenue for future research.

On the second characteristic, the SFNO-BVMC ensemble mean realistically degrades with lead time: it has a similar ensemble mean spectra as the IFS ensemble mean. Figure 12 shows that the ensemble means of SFNO-BVMC and IFS ENS similarly degrade in power after 24, 120, and 240 h. For z_{500}, there is a notable decline in power between lead times of 24 and 120 h. This sharp decline is due to the nonlinear error growth that characterizes forecasts at lead times of 3–5 d. On synoptic scales ($\sim$ 1000 km in space and 3–5 d in time), SFNO-BVMC's ensemble mean has a similar de-

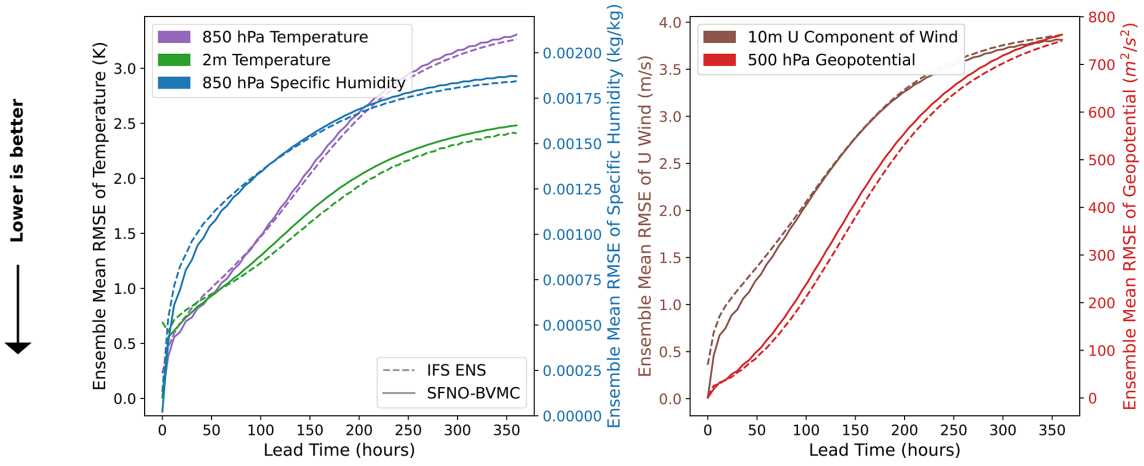


Figure 9. Ensemble mean RMSE of SFNO-BVMC and IFS ENS. SFNO-BVMC is the same 58-member ensemble described in Fig. 7. Scores are calculated from forecasts initialized at 732 initial conditions (two per day at 00:00 and 12:00 UTC) for 2020. SFNO-BVMC is validated against ERA5, and IFS ENS is validated against ECMWF's operational analysis. IFS ENS scores are taken from WeatherBench 2 (Rasp et al., 2024).

cline in power as IFS ENS. This increases our trust that the ensemble member trajectories realistically diverge, and the ensemble is correctly representing synoptic error growth.

These two results pass a crucial test laid out by Bonavita (2023). They originally posed this test comparing the spectra of a deterministic PanGu ML model and the IFS ensemble mean. Despite the blurring in PanGu, they show that a control run of PanGu does not successfully mimic the IFS ensemble mean spectrum. We present an ensemble prediction system from multiple deterministic ML models that has the above two characteristics.

3.3 Extreme diagnostics

The preceding analysis has evaluated ensemble weather forecasts from SFNO-BVMC on overall weather. This is necessary but as yet insufficient validation for our main scientific interest in LLHIs. Since extreme weather events are rare in space and time, they contribute relatively little to these scores. Hereafter, we focus on diagnostics specifically designed to validate the performance of SFNO-BVMC on extreme weather. We complement these diagnostics with a case study of the Phoenix 2023 heatwave in Fig. A1.

3.3.1 Extreme forecast index

As part of its IFS evaluation, ECMWF releases a Supplemental Score on Extremes (Haiden et al., 2023). This score is based on the extreme forecast index (EFI). Using an ensemble forecast and its associated model climatology, the EFI is a unitless quantity that quantifies how unusual an ensemble forecast is. The EFI ranges from -1 (unusually cold) to 1 (unusually hot). The EFI measures the distance between the ensemble forecast CDF and the model climatology CDF (Lalaurette, 2002; Zsótér, 2006). The formula for the EFI is

$$\mathrm{EFI} = \frac{2}{\pi} \int_0^1 \frac{Q - Qf(Q)}{Q(1-Q)} \mathrm{d}Q, \tag{2}$$

where Q is a percentile, and $Qf(Q)$ denotes the proportion of ensemble members lying below the Q percentile calculated from the model climatology. The model climatology is calculated for each lead time for each grid cell.

To calculate the EFI, a model climatology is necessary. The model climatology encapsulates the expected weather for a given time of year. For a given initial day, ECMWF creates a model climatology (called M-Climate) using hindcasts from nine initial dates per year, 20 years, and 11 ensemble members (for a total of 1980 values). The CDF of these 1980 values represents the model climatology. This CDF is defined at each grid cell for each lead time, and it is used to calculate the $Qf(Q)$ term in Eq. (2). See Lavers et al. (2016) for more information on the M-Climate definition.

We generate a model climatology of SFNO-BVMC using the same parameters as ECMWF's M-Climate, except the SFNO-BVMC M-Climate uses 12 ensemble members, not 11. This is due to the use of centered (positive and negative) bred vector perturbations, which requires an even number of ensemble members. After creating the climatology of SFNO-BVMC, we calculate the CDF of the model climate for each lead time for each grid cell. We use these CDFs to calculate the EFI for the SFNO-BVMC forecasts initialized on each day of summer 2023. Figure 13 visualizes a sample EFI from SFNO-BVMC and IFS 4 d into a forecast on an arbitrary summer day. The IFS EFI values are directly downloaded from the ECMWF data server. The SFNO-BVMC and IFS EFI values have excellent agreement across the globe (Fig. 13). Notable features include pronounced heatwaves

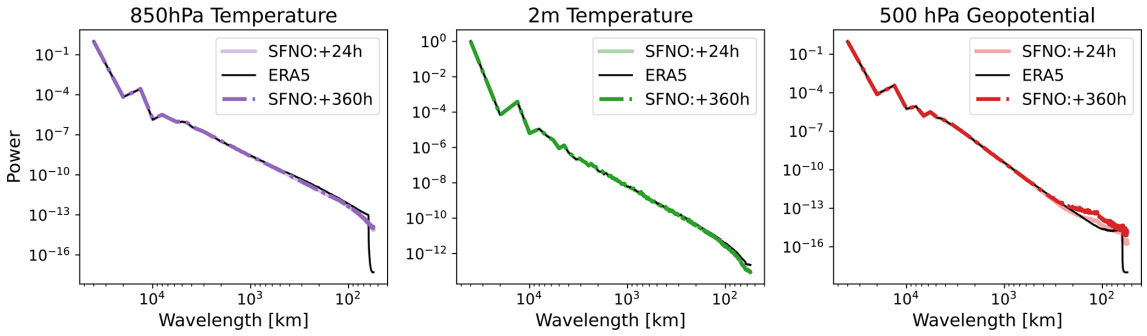


Figure 10. Control spectra. Spectra from the control member of SFNO-BVMC averaged across forecasts from 52 initial times, one per week starting 2 January 2020. Spectra are shown for 850 hPa temperature, 2 m temperature, and 500 hPa geopotential. Note the different scales on the y axis for each variable.

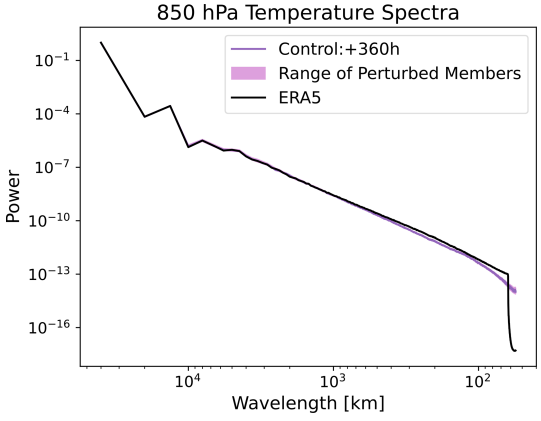


Figure 11. Perturbed spectra. Spectra of the control member and each perturbed member from a 58-member SFNO-BVMC ensemble are shown. The shading denotes the range of all the perturbed members. Spectra are averaged across forecasts from 52 initial times, one per week starting 2 January 2020.

over much of Africa, South America, and the Midwest of the United States. The strong El Niño pattern in the tropical Pacific appears in the EFI for both SFNO-BVMC and IFS ENS. Visually, SFNO-BVMC has a smoother EFI than IFS ENS. This is a consequence of the blurriness of the SFNO 2 m temperature predictions. Despite this, however, the SFNO EFI can still predict large-scale extremes, and the two models have similar scores on the extreme diagnostics below.

Figure 14 shows that IFS and SFNO-BVMC have highly correlated EFIs throughout summer 2023. Therefore, in principle, these two ensemble prediction systems offer comparable extreme forecasts and could be used to forecast various extreme events of interest. The EFI encapsulates the ability of each model to forecast extreme temperatures. Therefore, in principle, the EFI similarity between SFNO-BVMC and IFS means that they have similarly skillful extreme weather forecasts, including heat extremes and cold extremes of varying severity.

The EFI itself does not measure the accuracy of a forecast; it only measures how extreme or unusual a forecast is by comparing a given forecast to the model climatology. To evaluate the accuracy of the extreme forecast, the EFI is compared to an observational dataset to assess if the extreme forecasts match observations. We follow ECMWF's validation strategy of using a receiver operating characteristic curve to assess how well the EFI predicts the verification values.

3.3.2 Reliability and discrimination

Two key aspects of an ensemble forecast are its reliability and its discrimination. Measured by reliability diagrams, a forecast's reliability evaluates whether the predicted probability of extreme weather matches the observed occurrence. Measured by receiver operating characteristic (ROC) curves, forecast discrimination is the ability to distinguish between an extreme weather event and a non-extreme weather event. An ROC curve can be created for each forecast lead time, and it is summarized by the ROC area under curve (AUC) score. We calculated the ROC AUC for each lead time, and a purely random forecast would have an ROC AUC value of 0.5. A perfect forecast would have an ROC AUC value of 1.

Reliability diagrams and ROC curves are calculated by comparing two quantities: a binarized ground-truth value (1 or 0, for extreme and not extreme) and a continuous ensemble forecast between 1 and 0. A key validation criterion is the threshold defining extreme vs. not extreme. We calculate our threshold for extreme temperature using the same definition as Price et al. (2023). Using the years 1992–2016 of ERA5, we calculate the climatological 95th percentile 2 m temperature for each grid cell. These percentiles are calculated for each time of day (00:00, 06:00, 12:00, and 18:00 UTC) for each month. This results in 48 different thresholds in total. This definition of extreme accounts for the diurnal and seasonal cycles: an event is considered extreme if it is hot for the time of day and time of year. It thus includes warm nighttime temperatures, which have important implications for fire (Balch et al., 2022) and human health (Murage et al., 2017;

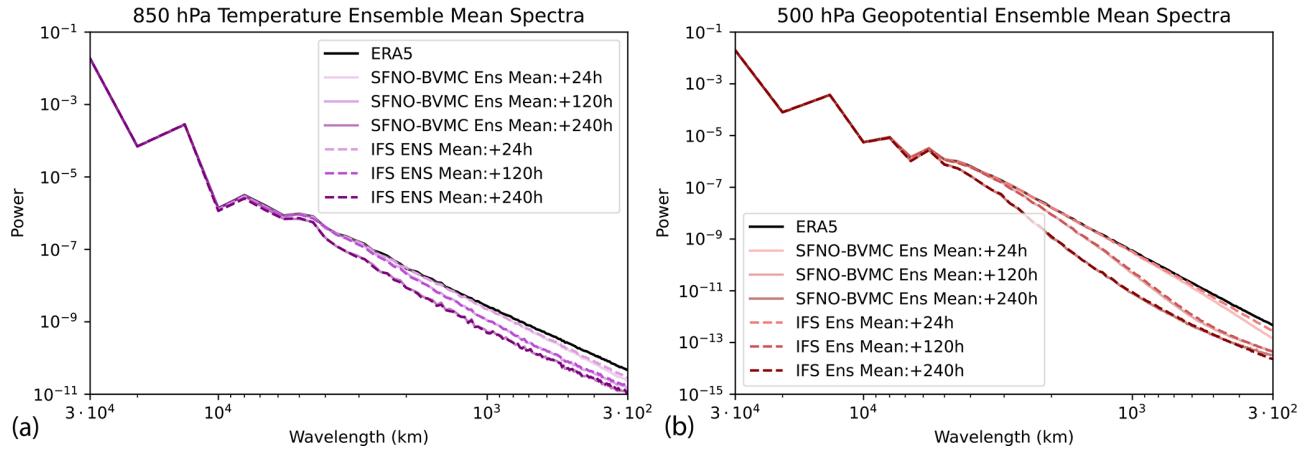


Figure 12. Ensemble mean spectra. The spectra of the ensemble mean of SFNO-BVMC and IFS ENS are shown. Spectra are averaged across forecasts from 52 initial times, one per week starting 2 January 2020. Spectra are shown for 850 hPa temperature **(a)** and 500 hPa geopotential **(b)**.

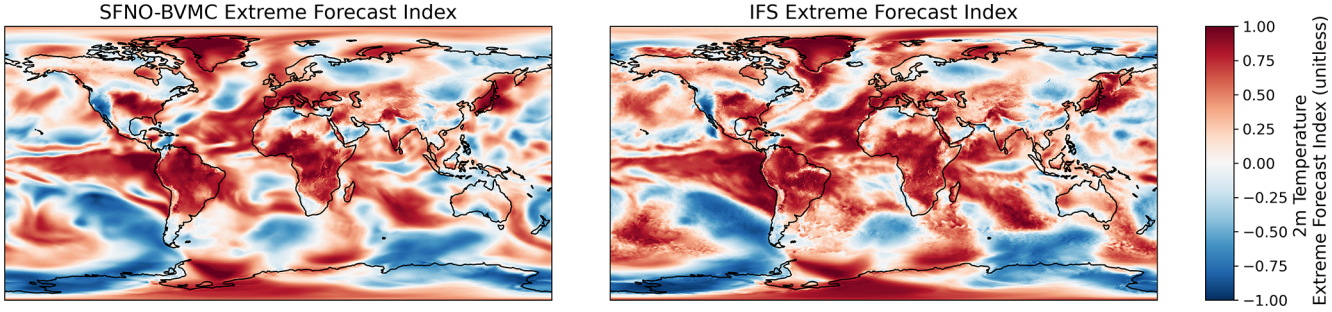


Figure 13. Visualization of the extreme forecast index from SFNO-BVMC and IFS ENS. For each grid cell and lead time, the extreme forecast index (EFI) is a unitless metric that represents the distance between the model climatology and the current ensemble forecast. It ranges from −1 (anomalously cold) to 1 (anomalously hot). For a sample 4 d forecast initialized on 19 August 2023, the EFI from the 58-member SFNO-BVMC is compared to the EFI from IFS ENS: the global latitude-weighted correlation is 0.89.

He et al., 2022), and warm winters, which have important implications for agriculture (Lu et al., 2022). This is a different rationale than defining extreme weather using an absolute temperature threshold or a threshold based only on the summer daily maximum.

Figure 15c shows that SFNO-BVMC and IFS ENS are similarly reliable in their prediction of extreme warm 2 m temperatures at lead times of 120 and 240 h. To create the reliability diagram in Fig. 15a, the ground-truth dataset is binarized using the extreme temperature threshold defined described above. The "forecast probability" is a continuous value from 0 to 1, indicating the proportion of the ensemble that exceeds the threshold. Over all grid cells and initial times of summer 2023, the reliability diagram compares the probabilistic forecasts of extreme events to their actual occurrence. In addition to the lead times in Fig. 15c, we visualize the reliability diagrams for other lead times (Fig. D1) and variables. We show that SFNO-BVMC also performs reliably when forecasting the heat index at lead times of 48, 96, 120, and 240 h. For 10 m wind speed and cold extremes, SFNO-BVMC matches the performance of the IFS ensemble (Figs. D2 and D3). However, we also show that at 240 h lead times, the model is not reliable when it confidently (greater than 50 % chance) forecasts wind extremes or cold temperature extremes (see Appendix D and Fig. D4 for more discussion). This is an area for future model development.

Next, we assess the ensemble's discrimination. Figure 15b shows that SFNO-BVMC and IFS ENS have a comparable ability to discriminate between extremes and non-extremes. Both ensembles have similar ROC area under curve (AUC) scores, which measure the discrimination of an ensemble. The ROC curve varies the threshold for classifying an event as "extreme" or "not extreme" from 0 to 1: for each threshold, the resulting true positive and false positive rates are plotted. A successful ROC curve would have a 0 false positive rate and 1 true positive rate: the area under such a curve would be 1. To calculate the ROC AUC scores in Fig. 15b, we use the EFI. To actually compare the EFI to observations, EFI ROC curves serve as ECMWF's Supplemental Score on Extremes in their IFS validation (Haiden et al., 2023). The

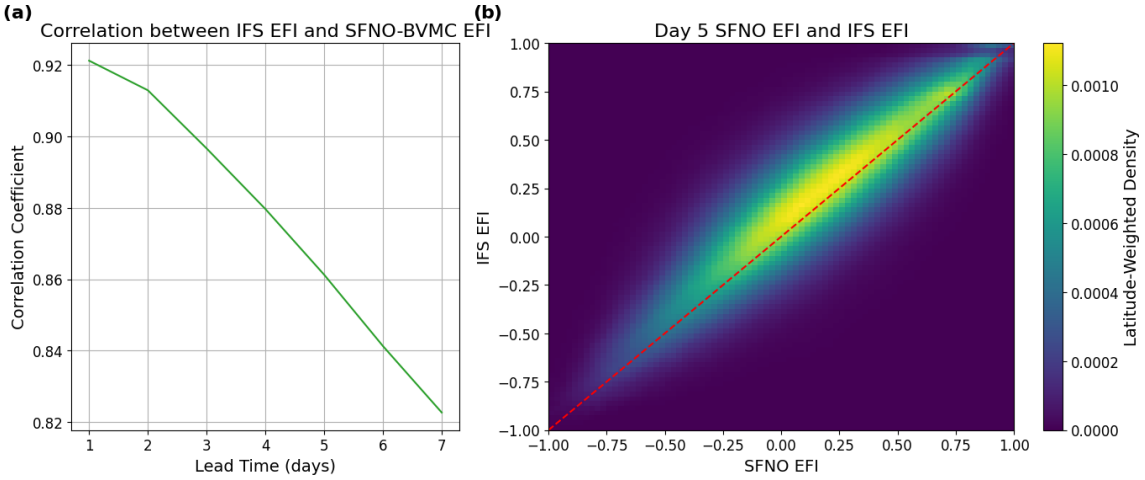


Figure 14. Comparing the SFNO-BVMC and IFS ENS extreme forecast index in boreal summer 2023. Panel **(a)** shows the latitude-weighted spatial correlation between IFS ENS EFI and SFNO-BVMC EFI as a function of lead time. Panel **(b)** shows the latitude-weighted 2D histogram between the SFNO-BVMC EFI and the IFS ENS EFI at a lead time of 5 d. Panels **(a)** and **(b)** are averaged using forecasts initialized over 92 initialization days, one per day (00:00 UTC) for each day in June, July, and August 2023.

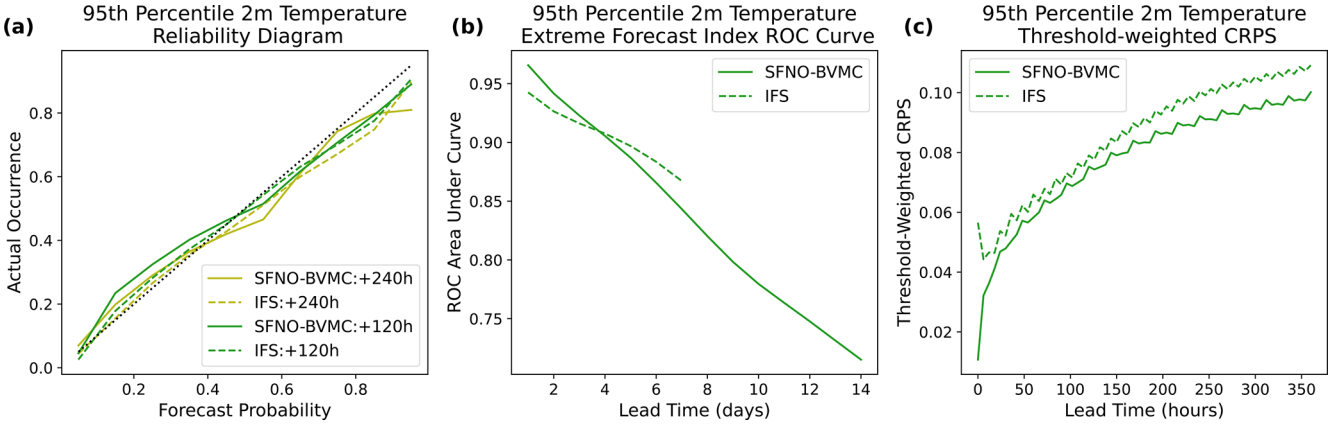


Figure 15. Extreme diagnostics of SFNO-BVMC and IFS ENS. Diagnostics are averaged over forecasts initialized at 00:00 UTC for each day in June, July, and August 2023 (total of 92 initialization days). SFNO-BVMC is validated against ERA5, and IFS ENS is validated against the ECMWF operational analysis. Panel **(a)** measures the receiver operating characteristic of the area under the curve. Higher is better. Panel **(b)** measures the threshold-weighted CRPS. Lower is better. Panel **(c)** measures the reliability diagram, which compares the forecast probability to the observed occurrence. Reliable ensemble forecasts appear along the one-to-one line.

IFS EFI is defined on a daily mean temperature, not a 6-hourly temperature. Therefore, the EFI ROC AUC score in Fig. 15b uses a threshold based on daily means. This results in 12 thresholds for extreme weather (one for each month), instead of 48 thresholds (one for each month for each time of day, as in Price et al., 2023). Based on the available data on the ECMWF MARS data server, we can only access IFS EFI values until a lead time of 7 d, so we only show IFS scores up to that lead time. At long lead times (approaching 14 d), much of the SFNO-BVMC EFI skill comes from the strong El Niño in summer 2023. Because the EFI is only calculated on data with a daily sampling frequency, Fig. 15b necessitated a different extreme threshold than Fig. 15a and c. This difference is necessary to enable comparison of EFI ROC curves with Haiden et al. (2023) and extreme diagnostics with Price et al. (2023).

3.3.3 Threshold-weighted continuous ranked probability score

We calculate threshold-weighted CRPS (twCRPS) on SFNO-BVMC and IFS ENS as a summary score. Since extreme weather events have tremendous societal consequences, a natural goal is to validate these weather forecasts specifically on their performance for such extremes. One approach might be to evaluate the forecasts during times of extreme weather. However, Lerch et al. (2017) explain the concept

of the forecaster's dilemma, which is a common pitfall that occurs with this strategy. This dilemma occurs when a forecast is validated on its extreme event forecasts only when those extremes actually happen. With this verification setup, a forecast system can hedge its performance by overpredicting extreme events. Since it would never be evaluated during common weather, the forecast would not be penalized for its overly extreme predictions. By construction, statistically proper scoring rules do not allow for such hedging, and twCRPS is one such scoring rule (Gneiting and Ranjan, 2011; Allen et al., 2023).

The equation for twCRPS is

$$\begin{aligned}\mathrm{twCRPS}(F, y, w) &= \int_{-\infty}^{\infty} (F(z) - 1\{y \le z\})^2 w(z)\mathrm{d}z \\ &= E_F|v(X) - v(y)| - \frac{1}{2}E_F|v(X) \\ &\quad - v(X')|, \end{aligned} \tag{3}$$

where w is a weighing function, X is a random variable drawn from the ensemble distribution, y is the verification value, and v is the antiderivative of w. We refer the reader to Allen et al. (2022) for further discussion of twCRPS and its derivation. We choose a weighing function

$$w = \begin{cases} 1 & \text{if } z > t \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This weighing function is applied at each grid cell. t is the 95th percentile 2 m temperature described above, calculated for each time of day for each month.

Equations (3) and (1) have the same structure; the difference is that Eq. (3) applies v to X and y. Therefore, twCRPS reduces to calculating the standard CRPS, when the ensemble and the ground truth are transformed using the following function (Allen et al., 2022):

$$\mathrm{twCRPS}(F, y) = \mathrm{CRPS}(F_t, \max(y, t)), \tag{5}$$

where F_t is the CDF of the transformed ensemble.

For each ensemble member E_i, where i goes from 1 to N for an ensemble size of N, the transformed ensemble member is

$$E_i' = \max(E_i, t). \tag{6}$$

The CDF of the transformed ensemble, $F_t(x)$, is thus calculated as $\frac{1}{N}\sum_{i=1}^{N} 1(\max(E_i, t) \le x)$ (Allen et al., 2022).

Similar to CRPS, twCRPS is calculated independently for each grid cell for each forecast initial time. After taking a global average and an average over each initial time in summer 2023, the twCRPSs are shown as a function of lead time in Fig. 15b.

twCRPS assigns no penalty when the ensemble forecast and the ground truth are below the extreme threshold. This is the most common situation that accounts for much of the CRPS, but it can mask out the performance on extremes. If an ensemble member lies above the threshold when the truth is below the threshold, then the ensemble will be penalized with a higher twCRPS. This is a solution to the forecaster's dilemma: the ensemble can no longer hedge its score by overpredicting extreme events above the threshold. If an ensemble forecast is below the threshold while the truth is above the threshold (false negative extreme), then the ensemble is also penalized. As the ensemble is transformed according to Eq. (6), this penalty is determined by the distance between the threshold and the truth, not the distance between the raw forecast and the truth. Therefore, twCRPS penalizes both overprediction and underprediction of extremes. It provides the benefits of the standard CRPS, as it evaluates a probabilistic forecast of a single ground-truth value.

Figure 15b shows that SFNO-BVMC and IFS have similar twCRPSs, with SFNO-BVMC performing slightly better on this metric. Since this score assesses the prediction of the tails of the distribution, SFNO-BVMC is a trustworthy model for predicting extreme 2 m temperature events. The twCRPS has the same units as the standard CRPS; for 2 m temperature, the unit is Kelvin. However, the values for twCRPS are lower than those for CPRS because the former assigns no penalty for the most common case when both the ensemble members and the ground-truth value are below the threshold. In those cases, the twCRPS will be 0. Because the ensemble members and the verification are transformed as in Eqs. (5) and (6), the twCRPS has a lower value than the CRPS.

twCRPS complements other forecast diagnostics, including those specifically focused on extremes. Recently, Ben Bouallègue et al. (2024) validated PanGu weather on extreme weather events, in part by comparing quantile–quantile plots of PanGu, IFS, and ERA5. While these plots compare the aggregate distributions of the forecasts and the truth, they do not assess whether extreme forecasts are collocated (in space and time) with extreme observations. Ben Bouallègue et al. (2024) state that additional diagnostic tools are necessary to evaluate this. We suggest that twCRPS fills this need, as it focuses on the tails of the ensemble distribution, but it also evaluates whether the forecasts coherently predict extremes at the right space and time.

4 Discussion and conclusion

In Part 1 of this two-part paper, we introduce SFNO-BVMC, an entirely ML-based ensemble weather forecasting system. This ensemble is orders-of-magnitude cheaper than physics models, such as IFS. It enables the creation of massive ensembles that can characterize the statistics of low-likelihood, high-impact extremes. Here, we present the ensemble design, which uses bred vectors as initial condition perturbations and multiple checkpoints as model perturbations. Multiple checkpoints are created by retraining SFNO from scratch,

with a different set of random weights when SFNO is first initialized. In this paper, we present a range of ensemble design choices and rationales for making these decisions; we list these in Table 1. To maximize dispersion, we use a large SFNO with a small scale factor and large embedding dimension, and we avoid multistep fine tuning.

We assess the fidelity of SFNO-BVMC on overall ensemble diagnostics, spectral diagnostics, and extreme diagnostics. This comprehensive pipeline is specifically designed for ensemble forecasts (not solely for deterministic ones). As the field of ML-based ensemble forecasting rapidly grows, we hope that other groups also use these statistics to evaluate their ensembles. On overall diagnostics, SFNO-BVMC's performance lags approximately 12–18 h behind IFS ENS. We present a pipeline to evaluate the ensemble's performance on extreme 2 m temperature, 10 m wind speed, and heat index events.

The spectral diagnostics demonstrate that individual ensemble members in the SFNO-BVMC have blurry predictions compared to ERA5. We minimize this as much as possible through a small scale factor, a large embedding dimension, and no autoregressive fine tuning. Still, some degree of blurring remains. However, our spectral diagnostics reveal that the spectra from SFNO-BVMC remain constant throughout the rollout. Additionally, the SFNO-BVMC ensemble mean spectra indicate that the ensemble members realistically diverge. Future research and architectural improvements are necessary to reduce the extent of the initial blurring.

Bred vectors are open-sourced through the `earth2mip` package, and they can readily be applied to other deterministic architectures. This enables out-of-the-box ensemble forecasts from the wide array of existing deterministic architectures. Indeed, recently, there have been over 20 deterministic ML weather prediction models (Arcomano et al., 2020; Bi et al., 2023; Nguyen et al., 2023; Chen et al., 2023b; Bodnar et al., 2024; Mitra and Ramavajjala, 2023; Ramavajjala, 2024; Pathak et al., 2022; Bonev et al., 2023; Weyn et al., 2021; Willard et al., 2024; Keisler, 2022; Karlbauer et al., 2023; Rasp et al., 2024, 2020; Lang et al., 2024; Couairon et al., 2024; Scher and Messori, 2021; Chen et al., 2023a). It is computationally expensive and programmer time-intensive to convert all these architectures into ensembles using probabilistic training (e.g., through diffusion models or through training on the CRPS loss function). Even for the architectures that are converted to probabilistic training, bred vectors and multiple checkpoints can provide baseline ensemble scores. This baseline can be used to guide further development of end-to-end training.

Understanding how ML models respond to perturbations is an important research frontier (Bülte et al., 2024; Selz and Craig, 2023). In particular, future work is necessary to compare the computational cost and skill of different initial condition perturbation methods (Bülte et al., 2024) in tandem with model perturbations. We find that bred vectors are a computationally inexpensive way to achieve reasonable spread–error ratios and to generate an arbitrarily large ensemble. Further refinement of initial condition perturbation techniques is needed to improve forecast performance. Two advantages of bred vectors are that they do not rely on external sources, and they can be used to generate arbitrarily large ensembles. First, Price et al. (2023) used external perturbations from operational data assimilation to include estimates of observational uncertainty. With the PanGu ML model, Bülte et al. (2024) tested ML ensembles with IFS perturbations. Bred vectors do not rely on external sources. If an ML model is used to emulate climate models (e.g., in Watt-Meyer et al., 2023), bred vector perturbations are still available, unlike IFS or data assimilation perturbations. Second, there are often a limited number of external perturbations from existing weather centers. Bred vectors can be used to generate arbitrarily large ensembles, such as the huge ensemble in Part 2.

Looking to the future of ML-based ensemble forecasting, an important design choice is whether the ensemble is created during training or after training. NeuralGCM (Kochkov et al., 2023) and GenCast (Price et al., 2023) create ensembles end to end during training; they train using probabilistic loss functions. Here, we train SFNO using a deterministic loss function, and we create the ensemble after training. In the machine learning literature, it is an active area of research whether ensemble training or post hoc ensembling leads to the most reliable results (Jeffares et al., 2023). In weather forecasting, so far, GenCast and NeuralGCM offer superior ensemble performance to SFNO-BVMC. They have better CRPSs and spread–skill ratios. Even at full ERA5 horizontal resolution, GenCast does not produce blurry forecasts. While GenCast and SFNO-BVMC run on different hardware (TPUs compared to NVIDIA GPUs used here), GenCast takes 6 min to create a 2-week forecast with a time step of 12 h. At the same horizontal resolution, SFNO-BVMC takes 1 min to create a 2-week forecast with a time step of 6 h; therefore, SFNO-BVMC appears to be a factor of 12 faster for inference. In part, this difference is because SFNO-BVMC does not require the iterative denoising used by GenCast at each time step. In Part 2 of this paper, we assess the performance of huge ensembles of SFNO-BVMC. A promising area of future research is to explore the behavior of huge ensembles from these other ML-based models.

The current generation of ML-based ensemble weather forecasts all have core design differences. IFS ENS uses physics-based modeling, NeuralGCM uses a differentiable dynamical core and an ML physics parameterization, GenCast uses a diffusion-based generative model, and SFNO-BVMC uses deterministic training. Because of these differences, future work is necessary to assess the strengths and weaknesses of each model in different meteorological regimes. When different forecasting systems have uncorrelated errors, a multimodel ensemble can lead to improved skill. Each forecasting system could be post-processed, bias-

corrected, and optimized to create the best ensembles for each region.

As the use of machine learning and huge ensembles grows in weather forecasting, it is important to consider climate equity (McGovern et al., 2024). Weather forecasts bring tremendous societal and economic value, and it is important to make them as accurate as possible globally (Linsenmeier and Shrader, 2023). Considerations of forecast skill should be improved for all locations, not just locations with large weather centers. One benefit of SFNO-BVMC is that it creates forecasts at a fraction of the computational cost. This means that organizations with limited access to large supercomputing resources can run weather forecasts and optimize them for their specific end use cases and datasets. In particular, they can be fine-tuned for regional purposes. In this introductory work, we primarily consider global metrics, and we focus on 2 m temperature. In the tropics, temperature variance is small due to a smaller Coriolis parameter, and humidity variations are particularly important, especially for impactful rainfall. Future work is necessary to consider the ensemble calibration and performance at the regional level, and this work can include explicit considerations of other variables, such as rainfall and humidity. In particular, the SFNOs trained here do not predict precipitation, and accurate medium-range rainfall forecasts are an important frontier in ML weather research.

In this paper, we run our extreme diagnostics pipeline on warm temperature extremes, and we validate on summer 2023, as it is the hottest summer in the observed record. At lead times of 48 and 96 h, the performance on cold extremes and wind extremes is similar to IFS. However, future work is necessary to reduce false positives for these other classes of extremes at 10 d lead times (Fig. D3). The EFI here is calculated on daily mean temperature, but it can also be calculated for other quantities, such as daily max or min temperature, convective available potential energy, or vapor transport (Lavers et al., 2016). Similarly, the ROC curves and reliability diagrams could be calculated for other types of extremes. We have presented an ensemble extreme diagnostics pipeline that can be used to guide development for other ML data-driven weather systems.

In Part 2, we use SFNO-BVMC to generate a huge ensemble with 7424 members. This ensemble is 150× larger than the ensembles used for operational weather forecasting. We explore how an ensemble of this size enables analysis of low-likelihood, high-impact extremes.

Appendix A: Case study: 2023 Phoenix heatwave

We include a case study for a heatwave in Phoenix in summer 2023. Phoenix had temperatures above 310 K (36.85 °C) for over 30 consecutive says in summer 2023. We compare the ensemble forecasts from SFNO-BVMC and the IFS ensemble in Fig. A1. We compare SFNO-BVMC to the IFS ensemble, and we visualize their respective verification datasets. We show that at a lead time of 3 d, SFNO-BVMC can forecast the high temperatures observed over the region during the heatwave. The IFS ensemble is initialized with an operational analysis, not ERA5, and we use this analysis as the verification dataset for IFS. Notably, the operational analysis has even sharper fields than ERA5: this has previously been quantified in Fig. S38 and Supplement Sect. 7.5.3 of Lam et al. (2023). The difference between operational analysis and ERA5 reanalysis is shown for this heatwave in Fig. A1.

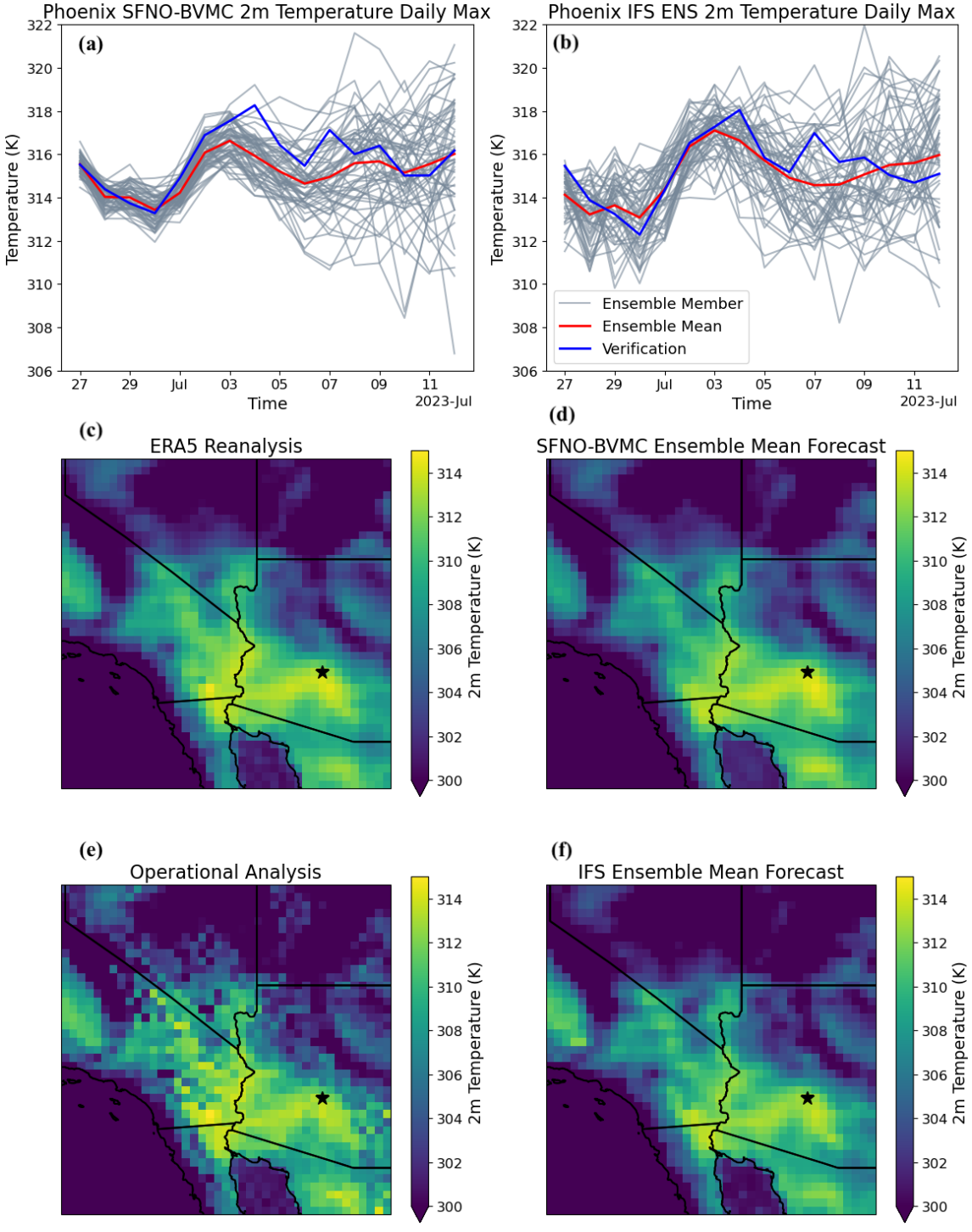
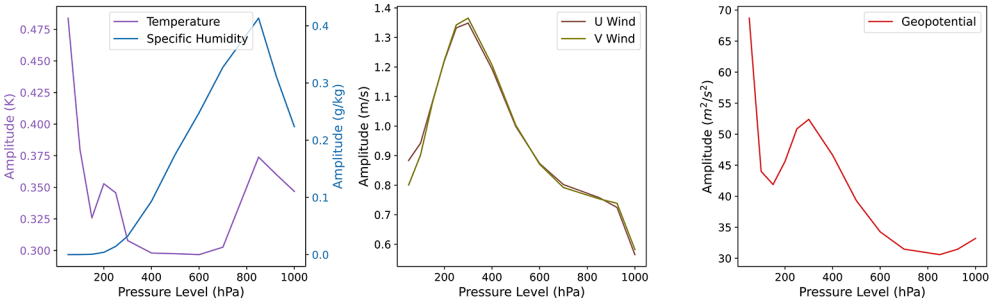


Figure A1. 2023 Phoenix heatwave. Comparison of ensemble forecasts from SFNO-BVMC **(a)** and the IFS ensemble **(b)** at a grid cell near Phoenix, Arizona, USA. Both models' forecasts are initialized on 27 June 2023 at 00:00 UTC. The SFNO-BVMC verification dataset is ERA5 and the IFS ENS verification dataset is operational analysis. Panels **(c)** and **(e)** show ERA5 and operational analysis, respectively, for the daily max temperature on 30 June 2023. Panels **(d)** and **(f)** show the SFNO-BVMC ensemble mean and IFS ENS mean, respectively, for the daily max temperature on 30 June 2023. The black stars in **(c)**–**(f)** denote the grid cell near Phoenix, Arizona, USA, used in **(a)** and **(b)**.

Appendix B: Perturbation amplitudes



Variable	Amplitude
u10m (m/s)	0.52
v10m (m/s)	0.53
u100m (m/s)	0.65
v100m (m/s)	0.66
t2m (K)	0.37
sp (Pa)	40.80
msl (Pa)	42.39
tcwv (mm)	0.95
d2m (K)	0.47

Figure B1. Bred vector perturbation amplitudes. The root mean square amplitude of the perturbation is shown for each variable.

The root mean square amplitude of the bred vector perturbations is set to be 0.35 times the deterministic RMSE of SFNO at 48 h. Figure B1 visualizes the actual numerical value of these amplitudes (with the factor of 0.35 applied).

Appendix C: Definition of spread and error

Below, we include our definitions for calculating the spread and error for calculation of the spread–error ratio (Fortin et al., 2014).

The ensemble forecasts have a 0.25 ° horizontal resolution on a regular latitude–longitude grid, so the ensemble forecasts have 721 latitude points and 1440 longitude points. Let i and j be the indices of a grid cell at a given latitude and longitude.

For each grid cell, the ensemble mean is

$$\mu(i,j) = \frac{1}{N}\sum_{n=1}^{N} x_n(i,j).$$

For each grid cell, the ensemble variance is

$$\sigma^2(i,j) = \frac{1}{N}\sum_{n=1}^{N} (x_n(i,j) - \mu(i,j))^2.$$

To calculate the spread in the spread–error ratio, we first calculate the ensemble variance at each grid cell. Then, we take the global latitude-weighted mean of this variance. Then, we take the mean over forecasts from multiple initial dates. Finally, we take the square root.

$$\text{Spread} = \sqrt{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{i=1}^{721}\sum_{j=1}^{1440} l(i,j)\sigma^2(i,j)},$$

where $l(i,j)$ denotes the latitude weight for grid cell i,j. The latitude weights enable calculation of the global mean.

We follow a similar process for calculating the error, except the ensemble variance is replaced with ensemble mean squared error.

$$\text{Error} = \sqrt{\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}}\sum_{i=1}^{721}\sum_{j=1}^{1440} l(i,j)(\mu(i,j) - y)^2},$$

where y denotes the verification value. The spread and error are calculated for each lead time and shown in Fig. 8. Results are shown for all forecasts in the test set year, 2020, so $\mathcal{T} = 732$, for 732 initial times (two per day).

Appendix D: Diagnostics for additional variables and lead times

We show the reliability of the forecasts from SFNO-BVMC at a lead time of 2 d and 4 d in Fig. D1. IFS is more reliable than SFNO, since its forecasts lie closer to the 1-to-1 line, though the performance is comparable. When SFNO-BVMC predicts 95th percentile temperature with approximately 20 % to 30 % probability, the actual occurrence is more frequent than this predicted probability.

We also include the reliability diagrams for the heat index (Lu and Romps, 2022), which combines 2 m temperature and moisture, and for 10 m wind speed. In Fig. D2, we show the reliability diagrams for 95th percentile heat index. As for 2 m air temperature, the 95th percentile heat index is calculated from 1993–2016 ERA5 climatology. Comparing Figs. D2 to 15, we find that SFNO-BVMC is similarly reliable for heat index extremes as it is for warm 2 m temperature extremes.

Next, we show overall CRPSs and reliability diagrams for 10 m wind speed, and we also show the model's reliability for forecasting cold extremes (5th percentile 2 m temperature). To calculate these statistics, we use December–January–February (DJF) from 2021 to 2022, while we use summer 2023 in Fig. 15. We use a winter season for the Northern Hemisphere to include sufficiently cold extremes over land in the Northern Hemisphere. Additionally, DJF 2021–2022 is within the time period included by the WeatherBench dataset (Rasp et al., 2024), so we can readily access

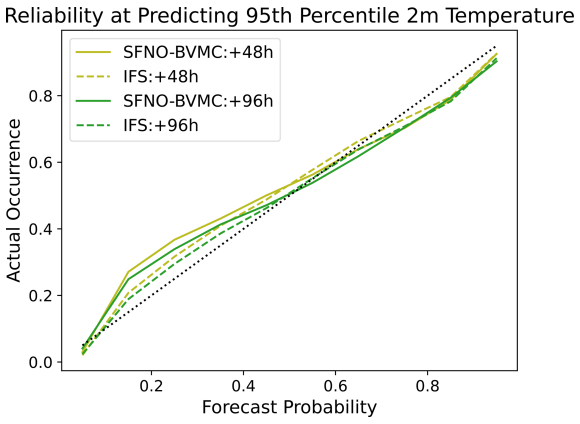


Figure D1. Reliability diagram at 48 and 96 h lead times. Reliability diagrams are shown for 95th percentile extremes at a lead time of 48 and 98 h. Reliability diagrams are calculated using all initial times from summer 2023. Successful forecasts lie along the 1-to-1 line.

the IFS ensemble wind forecasts via a Zarr file stored on Google Cloud, without having to download additional data from ECMWF's tape servers. For 10 m wind speed and cold extreme air temperature, SFNO-BVMC and the IFS ensemble have similar performance on overall CRPS and on reliability diagrams (Fig. D3). This indicates that the ensemble generation methodology (bred vectors and multiple checkpoints) is promising for other variables and classes of extremes. SFNO-BVMC does degrade in reliability when making forecasts of extreme wind with 90 % probability. This problem is accentuated at longer lead times (see below), and future research is necessary to isolate the cause of this behavior.

We identify two areas for future research and model improvement. First, interestingly, both SFNO-BVMC and the IFS ensemble have degraded performance for forecasting cold extremes, as opposed to warm extremes (compare the 5th percentile reliability in Fig. D3 to the 95th percentile reliability diagrams in Fig. D1). With future model development, we hope to improve the performance of SFNO-BVMC in forecasting cold extremes. Second, for 10 m wind speed and cold temperature extremes at a lead time of 10 d, SFNO-BVMC's reliability degrades (Fig. D4). For these variables, the ensemble still has a good overall forecast score (see the wind speed CRPS in Fig. D3 and the 2 m temperature Fig. 7). SFNO-BVMC reliability is close to IFS for extreme forecast probabilities from 0 % to 50 %. However, the reliability drops when the model predicts high probabilities (greater than 70 %) of extreme conditions. In these cases, SFNO-BVMC tends to be overconfident: its forecast of an extreme event does not match the observed outcome. This overconfidence occurs extremely rarely. At a lead time of 10 d, it is very uncommon (less than 1 % of all forecasts for 2 m temperature, less than 0.1 % for 10 m wind speed) for SFNO-BVMC to predict a greater than 70 % chance of extreme wind or cold temperatures. At this long lead time, there is significant ensemble spread induced by the perturbations, so the ensemble system is not confident in issuing extreme forecasts. Still, having a calibrated reliability diagram is crucial for all forecast probabilities, and this shortcoming must be resolved with future model development.

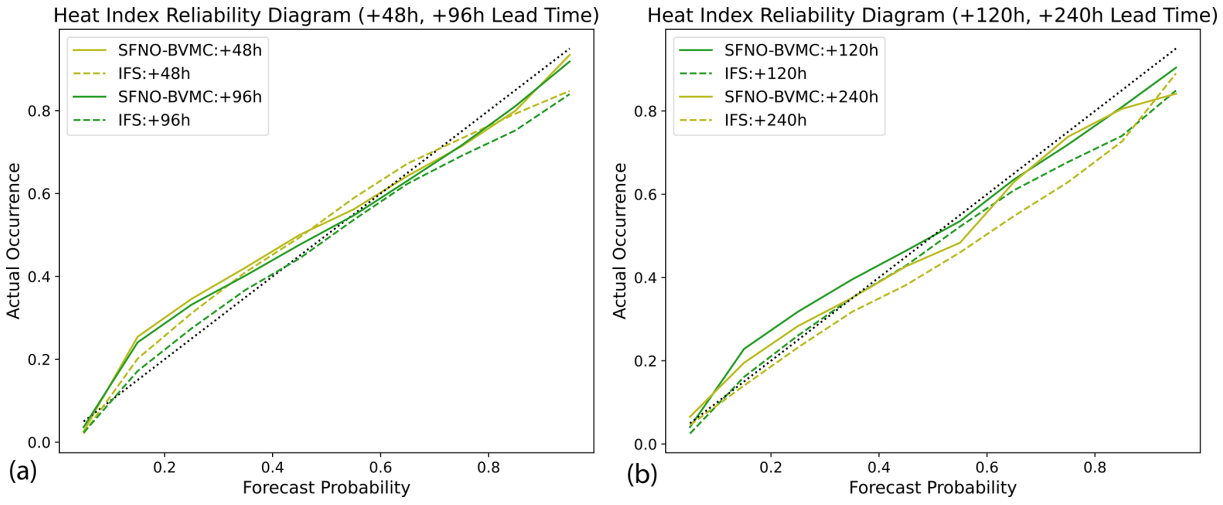


Figure D2. Reliability diagram for heat index. The heat index combines 2 m air temperature and 2 m dew point. Reliability diagrams are shown for 95th percentile heat index extremes at a lead time of 48 and 98 h **(a)** and 120 and 240 h **(b)**. Reliability diagrams are calculated using summer 2023 forecasts, from 1 June 2023 to 14 August 2023.

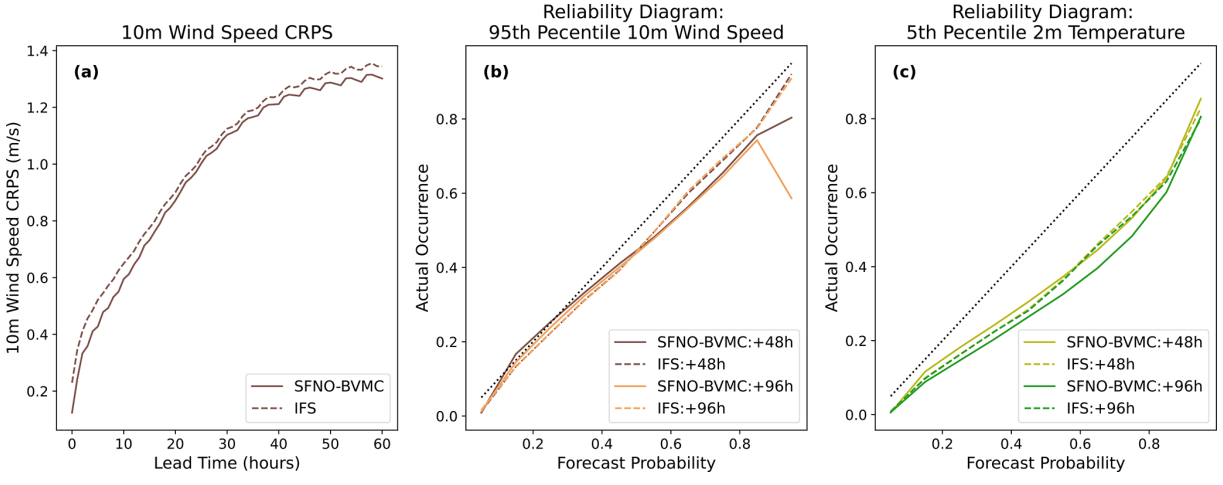


Figure D3. SFNO-BVMC performance on 10 m wind speed and cold extremes. **(a)** Overall CRPS for SFNO-BVMC and the IFS ensemble on 10 m wind speed. Lower CRPSs are better. **(b)** Reliability diagrams for 95th percentile 10 m wind speed for SFNO-BVMC and IFS at lead times of 48 and 96 h. **(c)** Reliability diagrams for 5th percentile temperature extremes for SFNO-BVMC and IFS at 48 and 96 h lead times. All scores are calculated using all forecasts initialized in December–January–February 2021. Successful forecasts lie along the 1-to-1 line.

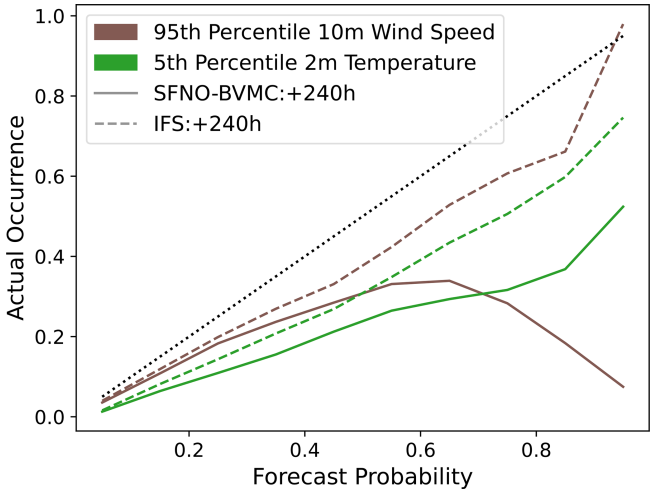


Figure D4. Reliability diagram at 240 h lead time for 95th percentile wind speed and 5th percentile temperature extremes. Reliability diagrams are shown for 95th percentile extremes at a lead time of 48 and 98 h. Reliability diagrams are calculated using all forecasts initialized in December–January–February 2021. Successful forecasts lie along the 1-to-1 line.

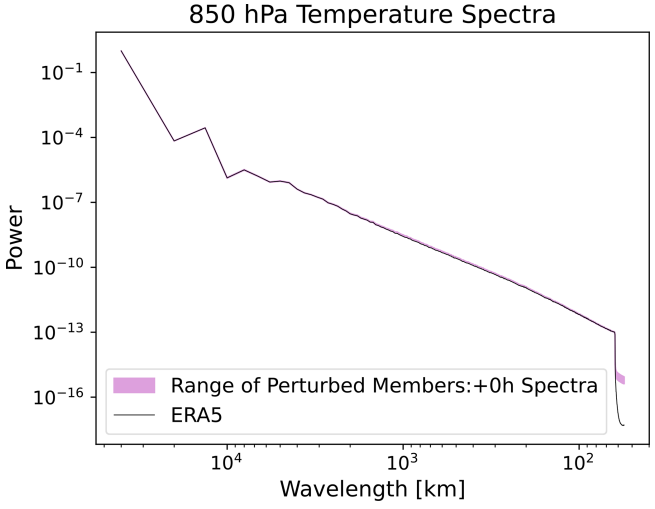


Figure D5. Perturbed spectra at 0 h. Same as Fig. 11, but showing the spectra of perturbations applied to the ERA5 initial conditions.

Appendix E: Solar zenith error in computer code when calculating initial condition perturbations

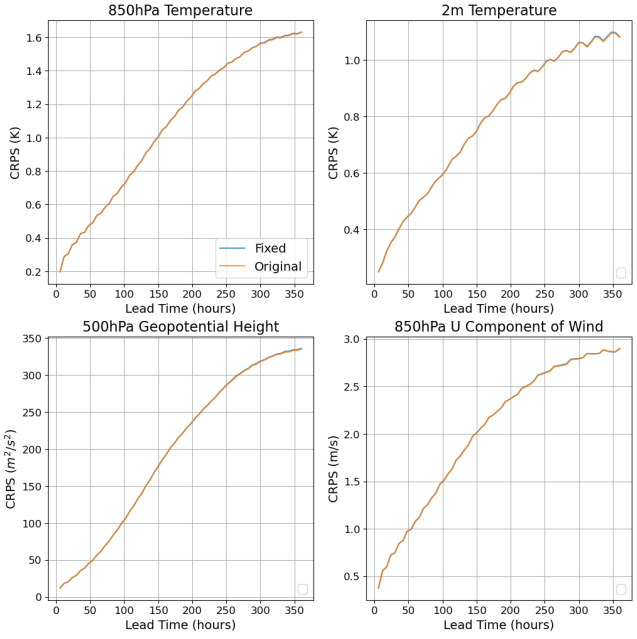


Figure E1. CRPS comparison of original and fixed bred vector perturbation method. Our original calculation of bred vectors contained an error with a mismatched cosine zenith angle during the first two breeding steps. This figure compares the CRPS of an ensemble with the "original" (incorrect) bred vector calculation and the "fixed" calculation for 2 m and 850 hPa atmospheric temperatures, the 500 hPa geopotential height, and the 850 hPa zonal wind as representative fields. Results are shown for 52 forecasts initialized in summer 2020, one per week starting 2 January.

After performing the analysis in this paper, we discovered an error in our calculation of the initial condition perturbations. During the first steps of calculating bred vectors at t_{-2} and t_{-1} (Fig. 4), we incorrectly supplied SFNO with the solar zenith angle at time t_0. We have verified and established that this error does not affect any of the conclusions or scores presented in this paper. This error does not make a discernible difference for three reasons. First, the last breeding step calculates the perturbations using SFNO at t_0. At t_0, the correct zenith angle is supplied, so the final perturbation is still based on the correct SFNO forecasts. Second, we confirm that the error does not cause undesired artifacts related to the diurnal cycle in the actual perturbations (see Fig. 5). Third, the breeding cycle only uses one-step forecasts, which means that the error from using the incorrect zenith angle does not grow.

We also note that this error does not affect the 15 d rollout of SFNO, only the calculation of the bred vectors. Given the nature of this error, we do not believe it would cause SFNO-BVMC to appear better than it actually is. Instead, it would be more likely to degrade its performance, making the method seem worse than it really is. We compare the ensemble scores for the ensemble with the error (named "original") and the fixed ensemble (named "fixed") in Figs. E1 and E2; the scores are nearly identical. We have corrected the error on the GitHub page for our project, but for scientific reproducibility, the error remains in the code base in the DOI in our "Code and data availability" section, since this is the version of the code that we used for our analysis.

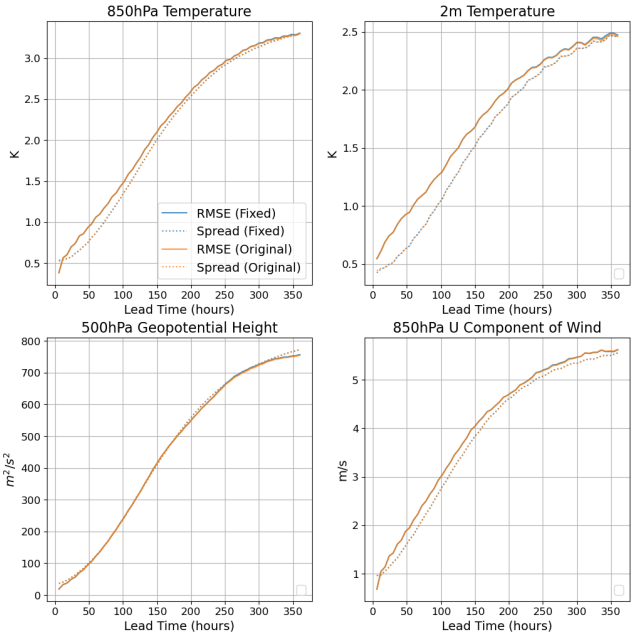


Figure E2. Ensemble mean RMSE and ensemble spread comparison of original and fixed bred vector perturbation method. Our original calculation of bred vectors contained an error with a mismatched cosine zenith angle during the first two breeding steps. This figure compares the ensemble mean RMSE and ensemble spread of an ensemble with the "original" (incorrect) bred vector calculation and the "fixed" calculation. Results are shown for 52 forecasts initialized in summer 2020, one per week starting 2 January.

Code and data availability. The code, datasets, and models used to produce the results used in this paper are archived on DataDryad under https://doi.org/10.5061/dryad.2rbnzs80n (Mahesh et al., 2025b). The code is integrated with Zenodo at the aforementioned DOI, and it is also available at https://github.com/ankurmahesh/earth2mip-fork (Mahesh et al., 2025c) as an additional download location. We include the code to train SFNO, conduct ensemble inference with bred vectors and multiple checkpoints, and scoring and analysis code. We also open-source the model weights of the trained SFNO. See the README of the DOI for information on how to use the code base and for the permission license associated with the code and data. The code is available via the Lawrence Berkeley Lab BSD variant license, and the data are available with a CC0 license. To run the ensemble for inference, a current version of the project is available from the project website at https://github.com/NVIDIA/earth2studio (NVIDIA, 2025) under the Apache-2.0 license.

Author contributions. AM and WDC contributed equally to this work. AM, BB, NB, JE, YC, PH, TK, JN, TAO, MR, DP, SS, and JW wrote software and performed formal data analysis. WDC, KK, and MP supervised the research project. WDC, KK, and MP acquired funding for the project. WDC, KK, PH, SS, AM, and MP obtained computational resources for the project. All authors contributed to the methodology of the project. WDC, AM, BB, YC, PH, KK, JN, TAO, MP, MR, SS, and JW contributed to the conceptualization of the project.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Geoscientific Model Development*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Special issue statement. This article is part of the special issue "Theoretical and computational aspects of ensemble design, implementation, and interpretation in climate science (ESD/GMD/NPG inter-journal SI)". It is not associated with a conference.

Acknowledgements. This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the US Department of Energy under contract no. DE-AC02-05CH11231 and by the Regional and Global Model Analysis Program area within the Earth and Environmental Systems Modeling Program. The research used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the US Department of Energy, under contract no. DE-AC02-05CH11231. The computation for this paper was supported in part by the DOE Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC) 2023–2024 award "Huge Ensembles of Weather Extremes using the Fourier Forecasting Neural Network" to William Collins (LBNL). This research was also supported in part by the Environmental Resilience Institute, funded by Indiana University's Prepared for Environmental Change Grand Challenge initiative.

Financial support. This research has been supported by Biological and Environmental Research (grant no. DE-AC02-05CH11231).

Review statement. This paper was edited by Po-Lun Ma and reviewed by Peter Düben and one anonymous referee.

References

Agrawal, S., Carver, R., Gazen, C., Maddy, E., Krasnopolsky, V., Bromberg, C., Ontiveros, Z., Russell, T., Hickey, J., and Boukabara, S.: A Machine Learning Outlook: Post-processing of Global Medium-range Forecasts, arXiv [preprint], https://doi.org/10.48550/ARXIV.2303.16301, 2023.

Allen, S., Ginsbourger, D., and Ziegel, J.: Evaluating forecasts for high-impact events using transformed kernel scores, arXiv [preprint], https://doi.org/10.48550/ARXIV.2202.12732, 2022.

Allen, S., Bhend, J., Martius, O., and Ziegel, J.: Weighted Verification Tools to Evaluate Univariate and Multivariate Probabilistic Forecasts for High-Impact Weather Events, Weather Forecast., 38, 499–516, https://doi.org/10.1175/waf-d-22-0161.1, 2023.

Arcomano, T., Szunyogh, I., Pathak, J., Wikner, A., Hunt, B. R., and Ott, E.: A Machine Learning-Based Global Atmospheric Forecast Model, Geophys. Res. Lett., 47, 9, https://doi.org/10.1029/2020gl087776, 2020.

Balch, J. K., Abatzoglou, J. T., Joseph, M. B., Koontz, M. J., Mahood, A. L., McGlinchy, J., Cattau, M. E., and Williams, A. P.: Warming weakens the night-time barrier to global fire, Nature, 602, 442–448, https://doi.org/10.1038/s41586-021-04325-1, 2022.

Baño-Medina, J., Sengupta, A., Watson-Parris, D., Hu, W., and Monache, L. D.: Towards calibrated ensembles of neural weather model forecasts, ESS Open Archive, https://doi.org/10.22541/essoar.171536034.43833039/v1, 2024.

Ben Bouallègue, Z., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, B. Am. Meteorol. Soc., 105, E864–E883, https://doi.org/10.1175/bams-d-23-0162.1, 2024.

Bercos-Hickey, E., O'Brien, T. A., Wehner, M. F., Zhang, L., Patricola, C. M., Huang, H., and Risser, M. D.: Anthropogenic Contributions to the 2021 Pacific Northwest Heatwave, Geophys. Res. Lett., 49, 23, https://doi.org/10.1029/2022gl099396, 2022.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., Gupta, J. K., Tambiratnam, K., Archibald, A., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: Aurora: A Foundation Model of the Atmosphere, arXiv [preprint], https://doi.org/10.48550/ARXIV.2405.13063, 2024.

Bonavita, M.: On some limitations of data-driven weather forecasting models, arXiv [preprint], https://doi.org/10.48550/ARXIV.2309.08473, 2023.

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., and Anandkumar, A.: Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere, arXiv [preprint], https://doi.org/10.48550/ARXIV.2306.03838, 2023.

Bonev, B., Kamenev, A., and Kurth, T.: Makani: Massively parallel training of machine-learning based weather and climate models, GitHub [code], https://github.com/NVIDIA/modulus-makani/tree/v0.1.0 (last access: 20 August 2025), 2024.

Brankovic, C., Palmer, T. N., Molteni, F., Tibaldi, S., and Cubasch, U.: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting, Q. J. Roy. Meteor. Soc., 116, 867–912, https://doi.org/10.1002/qj.49711649405, 1990.

Brenowitz, N. D., Cohen, Y., Pathak, J., Mahesh, A., Bonev, B., Kurth, T., Durran, D. R., Harrington, P., and Pritchard, M. S.: A Practical Probabilistic Benchmark for AI Weather Models, arXiv [preprint], https://doi.org/10.48550/ARXIV.2401.15305, 2024.

Bülte, C., Horat, N., Quinting, J., and Lerch, S.: Uncertainty quantification for data-driven weather models, arXiv [preprint], https://doi.org/10.48550/ARXIV.2403.13458, 2024.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W.: FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead, arXiv [preprint], https://doi.org/10.48550/ARXIV.2304.02948, 2023a.

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

Collins, W., Pritchard, M., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Mahesh, A., and Subramanian, S.: Huge Ensembles of Weather Extremes using the Fourier Forecasting Neural Network, EGU General Assembly 2024, Vienna, Austria, 14–19 Apr 2024, EGU24-4460, https://doi.org/10.5194/egusphere-egu24-4460, 2024.

Couairon, G., Lessig, C., Charantonis, A., and Monteleoni, C.: ArchesWeather: An efficient AI weather forecasting model at 1.5° resolution, arXiv [preprint], https://doi.org/10.48550/ARXIV.2405.14527, 2024.

ECMWF: IFS Documentation, https://www.ecmwf.int/en/publications/ifs-documentation, last access: 17 July 2024.

Esper, J., Torbenson, M., and Büntgen, U.: 2023 summer warmth unparalleled over the past 2,000 years, Nature, 631, 94–97, https://doi.org/10.1038/s41586-024-07512-y, 2024.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?, J. Hydrometeorol., 15, 1708–1713, https://doi.org/10.1175/jhm-d-14-0008.1, 2014.

Gneiting, T. and Ranjan, R.: Comparing Density Forecasts Using Threshold and Quantile-Weighted Scoring Rules, J. Bus. Econ. Stat., 29, 411–422, https://doi.org/10.1198/jbes.2010.08110, 2011.

Haiden, T., Janousek, M., Vitart, F., Bouallègue, Z. B., Ferranti, L., Prates, F., and Prates, F.: Evaluation of ECMWF forecasts, including the 2023 upgrade, European Centre for Medium Range Weather Forecasts Reading, UK, https://www.ecmwf.int/en/elibrary/81389-evaluation-ecmwf-forecasts-including-2023-upgrade (last access: 20 August 2025), 2023.

He, C., Kim, H., Hashizume, M., Lee, W., Honda, Y., Kim, S. E., Kinney, P. L., Schneider, A., Zhang, Y., Zhu, Y., Zhou, L., Chen, R., and Kan, H.: The effects of night-time warming on mortality burden under future climate change scenarios: a modelling study, The Lancet Planetary Health, 6, e648–e657, https://doi.org/10.1016/s2542-5196(22)00139-5, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Jeffares, A., Liu, T., Crabbé, J., and van der Schaar, M.: Joint Training of Deep Ensembles Fails Due to Learner Collusion, arXiv [preprint], https://doi.org/10.48550/ARXIV.2301.11323, 2023.

Karlbauer, M., Cresswell-Clay, N., Durran, D. R., Moreno, R. A., Kurth, T., Bonev, B., Brenowitz, N., and Butz, M. V.: Advancing Parsimonious Deep Learning Weather Prediction using the HEALPix Mesh, arXiv [preprint], https://doi.org/10.48550/ARXIV.2311.06253, 2023.

Keisler, R.: Forecasting Global Weather with Graph Neural Networks, arXiv [preprint], https://doi.org/10.48550/ARXIV.2202.07575, 2022.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Lottes, J., Rasp, S., Düben, P., Klöwer, M., Hatfield, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., and Hoyer, S.: Neural General Circulation Models, arXiv [preprint], https://doi.org/10.48550/ARXIV.2311.07222, 2023.

Kurth, T., Subramanian, S., Harrington, P., Pathak, J., Mardani, M., Hall, D., Miele, A., Kashinath, K., and Anandkumar, A.: FourCastNet: Accelerating Global High-Resolution Weather Forecasting Using Adaptive Fourier Neural Operators, in: Proceedings of the Platform for Advanced Scientific Computing Conference, Davos, Switzerland, 26–28 June 2023, PASC '23, ACM, https://doi.org/10.1145/3592979.3593412, 2023.

Lalaurette, F.: Early detection of abnormal weather using a probabilistic Extreme Forecast Index, European Center for Medium-range Weather Forecasting Technical Memoranda, 373, https://doi.org/10.21957/ehfunckhs, 2002.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., and Rabier, F.: AIFS - ECMWF's data-driven forecasting system, arXiv [preprint], https://doi.org/10.48550/ARXIV.2406.01465, 2024.

Lavers, D. A., Pappenberger, F., Richardson, D. S., and Zsoter, E.: ECMWF Extreme Forecast Index for water vapor transport: A forecast tool for atmospheric rivers and extreme precipitation, Geophys. Res. Lett., 43, 11852–11858, https://doi.org/10.1002/2016gl071320, 2016.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster's Dilemma: Extreme Events and Forecast Evaluation, Stat. Sci., 32, 1, https://doi.org/10.1214/16-sts588, 2017.

Leutbecher, M. and Palmer, T.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, 2008.

Li, L., Carver, R., Lopez-Gomez, I., Sha, F., and Anderson, J.: Generative emulation of weather forecast ensembles with diffusion models, Science Advances, 10, 13, https://doi.org/10.1126/sciadv.adk4489, 2024.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A.: Fourier Neural Operator for Parametric Partial Differential Equations, arXiv [preprint], https://doi.org/10.48550/ARXIV.2010.08895, 2020.

Linsenmeier, M., and Shrader, J.G.: Global inequalities in weather forecasts, Socarxiv [preprint], https://ideas.repec.org/p/osf/socarx/7e2jf.html (last access: 20 August 2025), 2023.

Liu, X., Saravanan, R., Fu, D., Chang, P., Patricola, C. M., and O'Brien, T. A.: How Do Climate Model Resolution and Atmospheric Moisture Affect the Simulation of Unprecedented Extreme Events Like the 2021 Western North American Heat Wave?, Geophys. Res. Lett., 51, e2024GL108160, https://doi.org/10.1029/2024gl108160, 2024.

Lu, X., O'Neill, C. M., Warner, S., Xiong, Q., Chen, X., Wells, R., and Penfield, S.: Winter warming post floral initiation delays flowering via bud dormancy activation and affects yield in a winter annual crop, P. Natl. Acad. Sci. USA, 119, 39, https://doi.org/10.1073/pnas.2204355119, 2022.

Lu, Y.-C. and Romps, D. M.: Extending the Heat Index, J. Appl. Meteorol. Clim., 61, 1367–1383, https://doi.org/10.1175/jamc-d-22-0021.1, 2022.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., O'Brien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge ensembles – Part 2: Properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators, Geosci. Model Dev., 18, 5605–5633, https://doi.org/10.5194/gmd-18-5605-2025, 2025a.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., O'Brien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge ensembles part I design of ensemble weather forecasts with spherical Fourier neural operators; Huge ensembles part II properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators, DRYAD [code, data set], https://doi.org/10.5061/DRYAD.2RBNZS80N, 2025b.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., O'Brien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge ensembles part I design of ensemble weather forecasts with spherical Fourier neural operators; Huge ensembles part II properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators, GitHub [code], https://github.com/ankurmahesh/earth2mip-fork (last access: 20 August 2025), 2025c.

McGovern, A., Bostrom, A., McGraw, M., Chase, R. J., Gagne, D. J., Ebert-Uphoff, I., Musgrave, K. D., and Schumacher, A.: Identifying and Categorizing Bias in AI/ML for Earth Sciences, B. Am. Meteorol. Soc., 105, E567–E583, https://doi.org/10.1175/bams-d-23-0196.1, 2024.

McKinnon, K. A. and Simpson, I. R.: How Unexpected Was the 2021 Pacific Northwest Heatwave?, Geophys. Res. Lett., 49, 18, https://doi.org/10.1029/2022gl100380, 2022.

Mitra, P. P. and Ramavajjala, V.: Learning to forecast diagnostic parameters using pre-trained weather embedding, arXiv [preprint], https://doi.org/10.48550/ARXIV.2312.00290, 2023.

Mittermaier, M. P.: A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites, Weather Forecast., 29, 185–204, https://doi.org/10.1175/waf-d-12-00075.1, 2014.

Murage, P., Hajat, S., and Kovats, R. S.: Effect of nighttime temperatures on cause and age-specific mortality in London, Environmental Epidemiology, 1, e005, https://doi.org/10.1097/ee9.0000000000000005, 2017.

Nguyen, T., Shah, R., Bansal, H., Arcomano, T., Madireddy, S., Maulik, R., Kotamarthi, V., Foster, I., and Grover, A.: Scaling transformer neural networks for skillful and reliable medium-range weather forecasting, arXiv [preprint], https://doi.org/10.48550/ARXIV.2312.03876, 2023.

NVIDIA: NVIDIA Earth2Studio, GitHub [code], https://github.com/NVIDIA/earth2studio (last access: 20 August 2025), 2025.

Olivetti, L. and Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather and GraphCast, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-1042, 2024.

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., and Smith, L.: Ensemble prediction: A pedagogical perspective, ECMWF Newsletter, 106, 10–17, 2006.

Pasche, O. C., Wider, J., Zhang, Z., Zscheischler, J., and Engelke, S.: Validating Deep Learning Weather Forecast Models on Recent High-Impact Extreme Events, arXiv [preprint], https://doi.org/10.48550/ARXIV.2404.17652, 2024.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anandkumar, A.: FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, arXiv [preprint], https://doi.org/10.48550/ARXIV.2202.11214, 2022.

Price, I., Sanchez-Gonzalez, A., Alet, F., Ewalds, T., El-Kadi, A., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: GenCast: Diffusion-based ensemble forecasting for medium-range weather, arXiv [preprint], https://doi.org/10.48550/ARXIV.2312.15796, 2023.

Ramavajjala, V.: HEAL-ViT: Vision Transformers on a spherical mesh for medium-range weather forecasting, arXiv [preprint], https://doi.org/10.48550/ARXIV.2403.17016, 2024.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting, J. Adv. Model. Earth Sy., 12, 11, https://doi.org/10.1029/2020ms002203, 2020.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, J. Adv. Model. Earth Sy., 16, 6, https://doi.org/10.1029/2023ms004019, 2024.

Scher, S. and Messori, G.: Ensemble Methods for Neural Network-Based Weather Forecasts, J. Adv. Model. Earth Sy., 13, 2, https://doi.org/10.1029/2020ms002331, 2021.

Selz, T. and Craig, G. C.: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?, Geophys. Res. Lett., 50, 20, https://doi.org/10.1029/2023gl105747, 2023.

Seneviratne, S., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Luca, A. D., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., éan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1513–1766, https://doi.org/10.1017/9781009157896.013, 2021.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NMC: The Generation of Perturbations, B. Am. Meteorol. Soc., 74, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:efantg>2.0.co;2, 1993.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NCEP and the Breeding Method, Mon. Weather Rev., 125, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:efanat>2.0.co;2, 1997.

Vargas Zeppetello, L. R., Raftery, A. E., and Battisti, D. S.: Probabilistic projections of increased heat stress driven by climate change, Communications Earth and Environment, 3, 1, https://doi.org/10.1038/s43247-022-00524-4, 2022.

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., Brenowitz, N. D., Kashinath, K., Pritchard, M. S., Bonev, B., Peters, M. E., and Bretherton, C. S.: ACE: A fast, skillful learned global atmospheric model for climate prediction, arXiv [preprint], https://doi.org/10.48550/ARXIV.2310.02074, 2023.

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models, J. Adv. Model. Earth Sy., 13, 7, https://doi.org/10.1029/2021ms002502, 2021.

Willard, J. D., Harrington, P., Subramanian, S., Mahesh, A., O'Brien, T. A., and Collins, W. D.: Analyzing and Exploring Training Recipes for Large-Scale Transformer-Based Weather Prediction, arXiv [preprint], https://doi.org/10.48550/ARXIV.2404.19630, 2024.

Zhang, L., Risser, M. D., Wehner, M. F., and O'Brien, T. A.: Leveraging Extremal Dependence to Better Characterize the 2021 Pacific Northwest Heatwave, J. Agr. Biol. Envir. St., 1–22, https://doi.org/10.1007/s13253-024-00636-8, 2024.

Zhong, X., Chen, L., Li, H., Feng, J., and Lu, B.: FuXi-ENS: A machine learning model for medium-range ensemble weather forecasting, arXiv [preprint], https://doi.org/10.48550/ARXIV.2405.05925, 2024.

Zsótér, E.: Recent developments in extreme weather forecasting, European Center for Medium-range Weather Forecasting Newsletter, 107, 8–17 pp., https://doi.org/10.21957/kl9821hnc7, 2006.