Towards calibrated ensembles of neural weather model forecasts

Jorge Baño-Medina¹, Agniv Sengupta², Duncan Watson-Parris³, Weiming Hu⁴, and Luca Delle Monache¹

¹University of California San Diego
²University of California, San Diego
³UC San Diego
⁴School of Integrated Sciences, James Madison University, Virginie, USA

May 10, 2024

Abstract

Neural Weather Models (NWM) are novel data-driven weather forecasting tools based on neural networks that have recently achieved comparable deterministic forecast skill to current operational approaches using significantly less real-time computational resources. The short inference times required by NWMs allow the generation of a large ensemble potentially providing benefits in quantifying the forecast uncertainty, particularly for extreme events, which is of critical importance for various socio-economic sectors. Here we propose a novel ensemble design for NWMs spanning two main sources of uncertainty: epistemic—or model uncertainty,— and aleatoric—or initial condition uncertainty. For the epistemic uncertainty, we propose an effective strategy for creating a diverse ensemble of NWMs that captures uncertainty in key model parameters. For the aleatoric, we explore the “breeding of growing modes” for the first time on NWMs, a technique traditionally used for operational numerical weather predictions as an estimate of the initial condition uncertainty. The combination of these two types of uncertainty produces an ensemble of NWM-based forecasts that is shown to improve upon benchmark probabilistic NWM and is competitive with the 51-member ensemble of the European Centre for Medium-Range Weather Forecasts based on the Integrated Forecasting System (IFS) —a gold standard in weather forecasting,— in terms of both error and calibration. In addition, we report better probabilistic skill than the IFS over land for two key variables: surface wind and air surface temperature.
Towards calibrated ensembles of neural weather model forecasts

J. Baño-Medina\textsuperscript{1}, A. Sengupta\textsuperscript{1}, D. Watson-Parris\textsuperscript{1,2}, W. Hu\textsuperscript{3}, L. Delle Monache\textsuperscript{1}

\textsuperscript{1}Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA
\textsuperscript{2}Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
\textsuperscript{3}School of Integrated Sciences, James Madison University, Harrisonburg, Virginia

Key Points:

• We present a strategy to produce ensemble forecasts with Neural Weather Models (NWMs), accounting for aleatoric and epistemic uncertainties.
• The proposed strategy reports better probabilistic skill than benchmark NWM and is competitive with the Integrated Forecasting System (IFS).
• We generate a 540-member ensemble in 4-5 hours using 1 GPU and in just a few minutes with 64 GPUs –orders of magnitude faster than the IFS.

Corresponding author: Jorge Baño-Medina, jbanomedina@ucsd.edu
Abstract
Neural Weather Models (NWM) are novel data-driven weather forecasting tools based on neural networks that have recently achieved comparable deterministic forecast skill to current operational approaches using significantly less real-time computational resources. The short inference times required by NWMs allow the generation of a large ensemble potentially providing benefits in quantifying the forecast uncertainty, particularly for extreme events, which is of critical importance for various socio-economic sectors. Here we propose a novel ensemble design for NWMs spanning two main sources of uncertainty: epistemic—or model uncertainty,— and aleatoric—or initial condition uncertainty. For the epistemic uncertainty, we propose an effective strategy for creating a diverse ensemble of NWMs that captures uncertainty in key model parameters. For the aleatoric, we explore the “breeding of growing modes” for the first time on NWMs, a technique traditionally used for operational numerical weather predictions as an estimate of the initial condition uncertainty. The combination of these two types of uncertainty produces an ensemble of NWM-based forecasts that is shown to improve upon benchmark probabilistic NWM and is competitive with the 51-member ensemble of the European Centre for Medium-Range Weather Forecasts based on the Integrated Forecasting System (IFS) —a gold standard in weather forecasting,— in terms of both error and calibration. In addition, we report better probabilistic skill than the IFS over land for two key variables: surface wind and air surface temperature.

Plain Language Summary
Numerical weather predictions (NWP) are prone to different sources of errors that propagate through time and space and grow exponentially due to the chaotic nature of the atmosphere. These aspects have triggered the pursuit of probabilistic forecasting in an effort to produce an ensemble of weather forecasts that allows for a description of possible outcomes given an initial weather state. Current operational tools based on NWP models can provide a limited number of members within the ensemble, due to the large computational requirements, which may impact their ability to capture a priori all likely scenarios. Recently, Neural Weather Models (NWM) based on neural networks have been able to provide high-quality weather forecasts at a fraction of the real-time computational cost of NWP-based systems. However, the mechanisms to capture the uncertainty of the prediction of these systems and deriving ensembles of forecasts remain unexplored. Here we develop a strategy to generate a 540-member ensemble that improves upon benchmark probabilistic NWM and is competitive with the 51-member ensemble of the European Centre for Medium-Range Weather Forecasts based on the Integrated Forecasting System —a gold standard in weather forecasting.

1 Introduction
Reliable and accurate weather forecasts are continuously demanded by practitioners in different socio-economic sectors (e.g., agriculture, tourism, water management) for reliable decision planning, to identify potential hazards or to allocate resources effectively, among others. Given an initial condition, global weather deterministic forecasts can be produced by a model — dynamical, statistical, or a hybrid combination,— that emulates the evolution of the atmosphere. Traditionally, forecasts have been based on Numerical Weather Prediction (NWP) models. NWP solves numerically a set of differential equations and execute sub-grid parameterization schemes to produce 4D multivariate, global atmospheric forecasts (Bauer et al., 2015). They require large computational resources to generate a forecast.

Neural Weather Models (NWM) are novel data-driven tools based on neural networks that require only a fraction of the resources demanded by NWP to generate a real-time prediction, and therefore might be potential cost-effective alternatives to generate
fast high-quality weather forecasts. Despite the short inference times reported (see e.g., Pathak et al. (2022)), they do require significant off-line resources to be trained. This training consists in learning a function describing the relationship between two atmospheric states separated in time (e.g., 6-hours), typically using global reanalysis data, such as the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5; Hersbach et al. (2020)). Once trained, given an initial condition they produce a forecast one time-step ahead. To generate forecasts farther into the future, we can repeat the operation $N$ times using as initial condition the prediction of the previous time-step (auto-regressive mode). Recently several topologies based on either graph neural networks (Lam et al., 2023) or transformers (Bi et al., 2023; Pathak et al., 2022) have reported global deterministic forecast skill on par with benchmark operational NWP models, such as the ECMWF Integrated Forecast System (IFS), and improved upon previous configurations exclusively based on convolutional neural networks (Rasp et al., 2020; Rasp & Thurey, 2021; Weyn et al., 2021; Scher & Messori, 2021).

Nevertheless, both NWP and NWM forecasts are prone to errors due to different sources of uncertainty in the system that can be broadly classified into two different categories: epistemic or model uncertainty, and aleatoric or initial condition uncertainty. Epistemic uncertainty arises from our lack of knowledge or deficiencies in the model formulation to represent the “true” evolution of the atmosphere. Aleatoric uncertainty evolves around errors in the initial condition due to, e.g., miss-calibration of atmospheric instrumentation, lack of data coverage, or approximation and assumptions in the data assimilation algorithms. In addition, the atmosphere is a chaotic system. This implies that even an infinitesimally small perturbation will grow in time, drastically altering the evolution of the atmosphere as compared to the non-perturbed trajectory (Lorenz, 1963; Selz & Craig, 2023). As a consequence, our ability to predict the future evolution of the atmosphere is limited in time.

These aspects have led to the development of probabilistic forecasting to generate an ensemble of weather forecasts that allow for a description of possible outcomes given an initial weather state (Leutbecher & Palmer, 2008). This can be achieved by slightly perturbing either the initial condition (aleatoric uncertainty) or the model (epistemic uncertainty), where each member then represents a different plausible evolution of the weather system. Ideally, the spread of a well-calibrated ensemble (i.e., measured by computing the standard deviation across members) will match the pattern of the errors, which can be estimated with the root mean squared error (RMSE) of the ensemble mean. To achieve the latter is far from trivial and in the last decades different ways have been explored to properly inject uncertainty into the weather prediction system and derive well-calibrated ensemble forecasts. Regarding the aleatoric uncertainty, one key condition is that the perturbations fields should represent the error patterns present in the analysis in both magnitude and shape. Decades of NWP research have transitioned from white noise perturbations to either Singular Vectors (SV, Molteni & Palmer (1993); Mureau et al. (1993)) or Breeding of Growing Modes (BGM, Toth & Kalnay (1997)). In the area of NWM simple methods have been tested so far to perturb the initial condition fields (e.g., based on white noise (Pathak et al., 2022), perlin noise (Bi et al., 2023), pink noise (Graubner et al., 2022) or lagged fields (Brenowitz et al., 2024)). To incorporate model uncertainty in the ensemble forecast, NWP builds on a variety of techniques that include utilizing different parameterization schemes (Palmer et al., 2009), stochastic physics perturbation (Sanchez et al., 2016) or multi-model ensembles (Medina et al., 2018), among others. Ultimately, these techniques aim to represent different physics in the atmosphere. In NWMs, model uncertainty can not be completely disentangled from the initial condition one, since models are ultimately trained using “imperfect” (re-)analysis data, and therefore they have also assimilated the latter source of error in the model coefficients. Some studies have performed model retraining to generate an ensemble of models. This methodology consists in starting from arbitrarily different initial configurations of coefficients in the optimization process (Scher & Messori, 2021). However, model retrain-
ing might be computationally demanding, since implies training from scratch every model of the ensemble. This goes against the main reason behind the exploration of NWM for weather forecasting, which is their computational benefits as opposed to NWP. A cost-effective alternative to model retraining is model sampling, where ensembles of models are built by selecting different sets of coefficients that correspond to different epochs in the optimization process (Weyn et al., 2021; Graubner et al., 2022).

Therefore, ensemble forecasting has been explored extensively for NWP, but the same is not true for NWM. There is a growing interest in the design of ensembles based on NWMs due to their ability to produce real-time forecasts with much shorter times than NWP, i.e., about 45,000 times faster than traditional NWP models on a node-hour basis (Pathak et al., 2022). This speedup could significantly increase the number of ensemble members generated and potentially revolutionize the field of probabilistic forecasting.

In this study, we propose an ensemble generation strategy based on both aleatoric and epistemic sources of uncertainty. For the aleatoric uncertainty, we examine the BGM approach for the first time on NWM. We select this method since it builds on current practices in NWP modeling—as opposed the ones tested so far by the community—and can be easily implemented with NWM. Meanwhile, for the epistemic, we propose an innovative strategy to build a diverse ensemble of NWM. We require 30 seconds to produce a single 7-day global weather forecast, which is orders of magnitude faster than the time demanded by ECMWF’s IFS predictive system. With a cluster of 64 P100 GPUs available, we are capable of generating a ~500-member ensemble in ~5 minutes, which is 10 times more members than ECMWF’s IFS ensemble.

2 Data

We use the ERA5 reanalysis to train the NWMs (Hersbach et al., 2020), which is a state-of-the-art dataset provided by the European Center for Medium-Range Weather Forecasts (ECMWF). Reanalysis datasets are combinations of observations and weather re-forecasts through data assimilation, and nowadays provide the most accurate representation of the atmosphere (Kalnay et al., 2018). The NWM models proposed so far in the observational space have used this dataset to train the models (Rasp et al., 2020; Rasp & Thuerey, 2021; Weyn et al., 2021; Scher & Messori, 2021; Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023). ERA5 provides hourly estimates for a large number of ocean-wave, land-surface and atmospheric variables at different pressure levels at 0.25° of spatial resolution, resulting into 780×1440 latitude-longitude fields. As a comparison to the NWMs, we also include in the analysis the 50-member ensemble from the ECMWF that builds on the IFS (see Section 3.2 for more details).

3 Methods

3.1 EnAFNO: Ensemble based on bred vectors and adaptive fourier neural operators

Figure 1 highlights the proposed strategy to generate ensemble of forecasts. This strategy builds on bred vectors for the aleatoric uncertainty and both model sampling and model retraining strategies for the epistemic uncertainty. We use the Adaptive Fourier Neural Operator (AFNO) deep learning topology or FourCastNet (Pathak et al., 2022) to predict a set of atmospheric variables 6-hours into the future. We used this model since the code was publicly available when we started working on this study and had memory requirements that fit in our computational infrastructure, but the strategy proposed could also be used to generate ensembles of forecasts with other NWMs topologies, such as Pangu-weather (Bi et al., 2023) or GraphCast (Lam et al., 2023). To produce forecasts at longer lead times, we perform inference in auto-regressive mode. The resulting
ensemble consists of 540 members in total, which is the result of perturbing the initial condition with 6 bred vectors and feed each of the perturbed fields to an ensemble of 90 AFNOs. We name this ENsemble generation strategy based on bred vectors and Adaptive Fourier Neural Operators: EnAFNO.

![Diagram of EnAFNO](image)

**Figure 1.** Schematic of the proposed strategy to generate an ensemble forecast with NWMs. An initial condition field, defined by the state of a set of atmospheric variables at time $t_0$, is perturb with 6 bred vectors (aleatoric uncertainty) and then feed into 90 distinct NWMs (epistemic uncertainty), a 540-member ensemble. Note that for illustration purposes we only show one variable in the initial condition field, but it actually builds on multi-variable input fields.

In the following sections we will go through the details of AFNO (Section 3.1.1), the breeding of growing modes method (Section 3.1.2), and the (proposed) strategy to quantify epistemic uncertainty of NWMs (Section 3.1.3). Finally, in Section 3.1.4 we describe the training parameters and computational requirements of EnAFNO.

### 3.1.1 Neural weather model: Adaptive Fourier Neural Operator (AFNO)

FourCastNet is a deep neural network (see e.g., Goodfellow et al. (2016) for a review in deep learning) that combines Fourier Neural Operators (FNO, Li et al. (2020)) with a Vision Transformer (ViT, Li et al. (2020)) backbone. ViTs have revolutionized a number of computer vision applications thanks to the multi-head self-attention mechanism that allows the network to extract complex features from the training data. The Adaptive FNO (AFNO) was first introduced by Guibas et al. (2021) as a cost-effective solution to perform multi-head self-attention in high-resolution spaces, since the potential number of combinations within a ViT scales with the number of features, turning them impractical to train. AFNO tackles the latter by performing self-attention in the Fourier domain. The features now become frequencies, consequently reducing the number of possible combinations.

AFNO was further adopted by Pathak et al. (2022) for weather forecasting as FourCastNet, becoming the first global data-driven NWM to operate at 0.25° of spatial resolution, which is comparable to global operational NWP models. FourCastNet forecasts the next 6 hours of weather defined by 20 key atmospheric variables. To produce long rollouts of forecasts, FourCastNet can be run in auto-regressive mode. One key modification to the original AFNO that is of relevance to this manuscript, is the training scheme (see Figure 3). The original FourCastNet training consists of two steps: pre-training and
fine-tuning. During pre-training (fine-tuning) the network minimizes the error between the forecasted fields 6 hours (6 and 12 hours) ahead and the ground-truth. Therefore, fine-tuning prepares the network to operate in auto-regressive mode. The authors of the study report training and inference times of 16 hours and 2 seconds (for a week-long forecast) on a cluster of 64 A100 GPUs, respectively. Therefore, inference is 45,000 times faster than traditional NWP models on a node-hour basis and requires 12,000 times less energy. In terms of forecast skill, FourCastNet shows comparable results to the IFS. The skill falls down considerably beyond weather time-scales into the sub-seasonal regime. In this regard, new research outlines the potential of incorporating the Earth’s geometry into the NWM (Bonev et al., 2023). We refer the reader to the original manuscripts for more details on the model (see Guibas et al. (2021) and Pathak et al. (2022)).

3.1.2 Aleatoric uncertainty: breeding of growing modes

Breeding of growing modes (BGM) was a technique proposed by Toth & Kalnay (1997) to estimate initial perturbations that “reflect the analysis errors of the day” in the data. The errors typically appearing in modern (re-)analysis datasets, which are combinations of forecasts (“first guess”) and observations, can be classified into two different types: growing (e.g., baroclinically unstable modes) and non-growing (e.g., gravity waves). The “first guess” forecast favours the propagation of growing errors, which are maintained and keep breeding through the successive combinations of “first guess” forecasts and observations over time (analysis cycle). As a consequence, growing modes usually dominate the error pattern. BGM aims to estimate the uncertainty on the data by simulating how the growing errors are propagated in the analysis cycle. The breeding cycle consists of 6 different steps (see Figure 2): 1) initialize a random white noise perturbation and add it to the weather state at time $t_o = t$; 2) in parallel, feed the perturbed and (3) non-perturbed atmospheric fields to the NWM; 4) compute the differences between the perturbed and non-perturbed forecasts: (5) scale back the differences to the size of the initial perturbation; (6) repeat steps 1-to-5 but using the bred vector as a perturbation to the weather state at $t_1 = t_o + \Delta t$. The vector of (scaled) differences is usually referred to as the bred vector. Therefore, bred vectors are representations of the fastest growing modes of error, which are obtained by redistributing the initial white noise field. We can compute several bred vectors by initializing the cycle again with different random seeds. Note that the bred vectors are NWM-dependent, and thus we have to go through the breeding cycle for each NWM independently.

3.1.3 Epistemic uncertainty: model retraining and model sampling

The machine learning community has tested different strategies to estimate the epistemic uncertainty. For neural networks, the majority of these methods were traditionally based on Bayesian statistics, where prior distributions are placed upon the NN parameters (Neal, 2012). Then the posterior distribution is computed given the training data, ultimately providing an estimate of the predictive uncertainty. However, in modern neural networks we very often find millions of parameters, such as in the skillful NWM, and thus inferring the posterior distribution in these cases is computationally intractable. One solution is to approximate the posterior using dropout, a recurrent element in deep learning topologies that randomly activates/deactivates neurons within the network (Gal & Ghahramani, 2016). Another solution is model retraining, where ensembles of neural networks are trained by stochastically tuning different elements during training, e.g., initial model parameters, training samples. This approach has been shown to be very effective in many applications (Lakshminarayanan et al., 2017), including weather forecasting (Weyn et al., 2021; Scher & Messori, 2021). In particular, Scher & Messori (2021) proved that model retraining of CNNs produced the most diverse ensemble of forecasts as compared to Bayesian dropout, white noise and singular vectors. Nevertheless, this approach is extremely computationally demanding for modern NWM, since training a
**Figure 2.** Schematic of the breeding cycle. At time $t_0$ an arbitrary perturbation (e.g., white noise) is added to the input atmospheric variables. In parallel, both the perturbed and non-perturbed analysis fields feed a neural weather model (e.g., \textit{AFNOx}). The difference between the control and perturbed forecasts is scaled back yielding the bred vector at time $t + ΔT$. This process is subsequently repeated by perturbing the analysis with the corresponding bred-mode.

A single model may require up to 10 days depending on the computational infrastructure—i.e., number and type of GPUs. For this reason, a few studies have performed \textit{model sampling}, which is based on the Stochastic Gradient Descent (SGD) iterations. SGD is the training algorithm used in neural networks to optimize the coefficients from an initial random state (Figure 3a) until convergence (Figure 3b), given a loss function. Therefore, the SDG iterations are repeated steps towards the direction of steepest descent in the loss surface, hence changing the initial model parameters. The size of the step is controlled by the learning rate, which is a tuneable parameter. An analysis of the model parameters at every epoch, suggests that the SGD trajectory contains useful information about the geometry of the posterior (Stephan et al., 2017; Maddox et al., 2019). Therefore, this methodology allows for efficient gathering of a large number of NWM by training only a single model from scratch and sampling models after convergence at different epochs (Weyn et al., 2021; Graubner et al., 2022). To maximize the diversity of the NWMs sampled, Stephan et al. (2017) suggested to place a high learning rate during this phase to enable a better exploration of the space of coefficients. One key drawback of this approach is that the models sampled not necessarily end up being diverse since SGD might simply oscillate towards a local/global minima (Figure 3d).

Here we propose a methodology to perform model sampling (Figure 3c), that builds on the 2-step training scheme that NWMs usually present: pre-training and fine-tuning. During pre-training (fine-tuning) the network minimizes the error between the forecasted fields in the next time step (two time steps) and the ground-truth. Therefore the loss function changes between training steps since the optimization objectives are different.
Figure 3. Schematic of the model sampling strategies in the context of FourCastNet’s training sequence. The dashed horizontal line at the top illustrates the two training steps—pre-training and fine-tuning,—described in Pathak et al. (2022). During pre-training (fine-tuning) the network minimizes the error between the forecasted fields 6 hours (6 and 12 hours) ahead and the ground-truth. Below, two horizontal parallel solid lines represent the evolution of NWMs (top) through the epochs (bottom), respectively. Each of the NWMs—or AFNOs in our case,—is represented by two sub-index, e.g., AFNOz_i, where z and i are the branch and model number within the branch, respectively (see Section 3.1.4). In (purple) pink we indicate the (proposed) benchmark model sampling strategies over the epochs where sampling occurs. Note that the color gets lighter with increasing epoch for both strategies to illustrate that usually a sufficient number is enough to approximate the posterior over the coefficients. Panels a-d are illustrations of hypothetical 2D error surfaces that represent different situations during training. Figure 3.a illustrates the initialization of the network coefficients (epoch 0) while Figure 3.b illustrates model convergence (epoch 150). Figures 3.c and 3.d are illustrations of the area explored by the proposed and benchmark model sampling strategies within the loss surface. Finally, the star is a representation of where the NWM is located at epochs 0, 150, 150 and 200 on panels a, b, c, and d, respectively.

We propose to start sampling models after changing the loss function (at the beginning of the fine-tuning step). We hypothesize that this approach allows us to sample a diverse number of NWMs by exploring a large area of the space of solutions, since we are no longer trapped in a local/global minima as we are training in a new loss surface (Figure 3c). We demonstrate this in Section 4.3.
3.1.4 Configuration of EnAFNO and computational requirements

We follow Pathak et al. (2022) and build on the following set of 20 variables from ERA5 at 00, 06, 12, 18 UTC: zonal and meridional wind velocities at 500, 850, 1000 hPa; geopotential height at 50, 500, 850, 1000 hPa; air temperature at 500 and 850 hPa; relative humidity at 500 and 850 hPa; zonal and meridional wind surface velocities, air surface temperature, surface pressure, mean sea level pressure and total column of water vapour. We scale the variables by subtracting the mean and dividing by the standard deviation over the training period. We follow the 2-step training procedure (i.e., pre-training and fine-tuning) described in Pathak et al. (2022). We use an Adam optimizer during the pre-training (fine-tuning) phase with a learning rate of 5E-4 and a cosine annealing scheduler (0.1 without scheduler) for 150 (35) epochs.

We inject epistemic uncertainty into the system by both model retraining and model sampling strategies. For model retraining, we train 3 branches of models, which we name AFNO-A, AFNO-B and AFNO-C. They are trained based on different seeds in the initialization of the model parameters and also different training periods: 1979-2015, 1979-2015 and both 1979-2015 and 2019-2022 for AFNO-A, AFNO-B and AFNO-C, respectively. For model sampling, we follow the strategy proposed in Section 3.1.3, and start sampling models during the fine-tuning step, beyond the epoch 156. Therefore, we leave a spin-up of 5 epochs to let the model adapt to the new error surface. In total we sample 30 models per branch since we observe no added value to the calibration skill beyond this quantity. We use the following nomenclature to refer to each of the models: $\text{AFNO}_{ij}$; where $i$ and $j$ represent the branch and number of sampled model within the branch, respectively (e.g., $\text{AFNO}_{A1}$).

To inject aleatoric uncertainty into the system we compute 3 independent breeding cycles per NWM (i.e., $\text{AFNO}_{ij}$). We initialize each of the cycles on the January 1st of 2018 with white noise of 0 mean and 0.15 of standard deviation. We let the cycle breed and maintain growing errors up to the 1st of March of 2018. For each breeding cycle we can then produce a pair of perturbations by either adding or subtracting the bred vector at time $t_0 = t$ to the corresponding (scaled) initial condition. Therefore, we can derive a total of 6 different members from the 3 independent breeding cycles per NWM.

Thus, the proposed ensemble of forecasts builds on 90 AFNO and 6 bred perturbations per NWM to derive a total of 540 members. This is almost 12 times more members than the IFS. Both the number of NWM models and bred vectors have been determined by testing different configurations (see Figure A1) and is further discussed in Sections 4.3 and 4.2. To train the models we used a cluster of 64 P100 GPUs. We report times of $\sim$7 days for training each branch, i.e., $\sim$21 days to derive the whole set of 90 AFNO, while to infer a 5-day 540-member global forecast on the same infrastructure takes $\sim$10 minutes. We note that shorter training/inference times than the ones presented here have been reported using better GPUs, such as the A100 (Pathak et al., 2022) or V100 (Bonev et al., 2023) type.

For illustration purposes, we show in Figure 4 the evolution of an atmospheric river over the Western coast of North America. We focus on this example, since it was already analyzed in the original FourCastNet study (Pathak et al., 2022) and is not included in the training data. Figure 4.a represents the total column water vapour for lead times of (from left to right) 6, 36, and 72 hours for (top to bottom) ERA5, and three random members of EnAFNO. We observe that EnAFNO is able to capture the main characteristics of the event showing consistent spatial patterns to those appearing in ERA5. Moreover, there are some differences among the three members shown. For instance, members 1 and 740 forecast larger total column water vapour values for the AR over the Pacific Ocean as compared to member 230 for a 36-hour of lead time. This exemplifies the variability of the NWM ensemble, that is able to represent different plausible evolutions of the atmospheric river. To visualize the performance of the entire ensemble, we average the to-
Figure 4. (a) Evolution of an atmospheric river over the Western coast of North America as represented by the total column water vapour. Model forecasts are initialized at 00:00 UTC on April 4, 2018. Example forecasts (from left to right) are shown for lead times 6, 36 and 72 hours and for ERA5 (row 1) and three random members of EnAFNO (rows 2-4); (b) Temporal series of the total column water vapour averaged over the Feather River Basin — latitudes \{39°, 41.5°\} and longitudes \{238°, 240.5°\}, — for the 540 EnAFNO members (turquoise), the ensemble mean (blue) and ERA5 (black).
3.2 Benchmark methods

We use the following benchmark models as comparison to EnAFNO:

- **IFS.** This ensemble builds on the ECMWF’s IFS (hereafter simply referred to as IFS). IFS consists of 50 members based on the Ensemble of Data Assimilation (EDA) system (Isaksen et al., 2010), that accounts for both model and initial condition uncertainty. It is considered to be the state-of-the-art NWP model in probabilistic forecasting. The forecasts are initialized every day at both 00:00 and 12:00 UTC, and have typically operated with spatial resolutions comparable to ERA5, what allows a direct comparison with the NWMs developed herein.

- **G-AFNO.** This ensemble was introduced in the original FourCastNet study as a first approximation to build ensemble of forecasts (Pathak et al., 2022). G-AFNO consists of 100 members, which are produced by perturbing the initial condition with 100 samples of a Gaussian distribution with 0 mean and 0.3 of standard deviation. These perturbed fields then feed a single AFNO, which has been trained following the guidelines described in Pathak et al. (2022). Therefore, this strategy only injects aleatoric uncertainty into the system. This ensemble represents the current benchmark involving the AFNO model. Since EnAFNO consists of 540 members, for a fair comparison we tested the sensitivity of the number of white-noise perturbations to the total spread of the ensemble, with no major differences beyond a 100-member ensemble (see Figure B1).

4 Results

We focus on four key atmospheric variables for different socio-economic sectors: the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa and the total column water vapour. We split the analysis into 3 different parts. First, we validate EnAFNO in terms of forecast skill and spread-error calibration (Section 4.1). Second, we evaluate the aleatoric uncertainty strategy of EnAFNO, i.e., the BGM method (Section 4.2). Third, we evaluate the proposed epistemic uncertainty strategy for EnAFNO and compare it against the approaches in prior literature (Section 4.3). We build on the following set of metrics (see Appendix C) to analyze the forecasts: standard error of the ensemble mean (skill), spatial ensemble spread (SES), binned spread-skill plots, and the Continuous Ranked Probabilistic Score (CRPS). We initialize the forecasts every day at 0:00 UTC from January 10th to the 28th of February of 2018. January 1st to January 10th are used to spin-up the bred vectors.

4.1 Evaluation of EnAFNO

We start by evaluating both the model error and model calibration of EnAFNO on a global scale. To this aim, Figure 5 displays the spread-skill plots for (in rows) the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa, and the total column water vapour at (in columns) different lead times: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). In colours we represent the 3 ensemble strategies: EnAFNO (blue), ENS (red), and G-AFNO (gray). The diagonal line represents the space of “perfect” calibration where the ensemble spread and the error are equal. If an ensemble has a higher (lower) error than spread this means that it is underdispersive (overdispersive), i.e., the probabilistic forecast is overconfident (underconfident). Please refer to Appendix C3 for details on the computation of these spread-skill plots. We observe that both the error and the spread of the forecasts increase as a function of lead time, regardless of the ensemble strategy considered. Nevertheless, the skill-spread ratio is different across variables, methods, and lead times. G-AFNO produces underdispersive ensembles, as already outlined in Graubner et al. (2022), and presents higher errors compared to other methods. Meanwhile, EnAFNO shows spread-skill val-
Figure 5. Binned spread-skill diagrams for the different ensemble strategies. Displayed by rows are the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa, and the total column water vapour, whereas the different lead times are displayed by columns: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). The colored lines represent the proposed ensemble EnAFNO (red), the IFS (blue) and the two benchmark probabilistic NWM (gray and green). We build ensemble forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in parenthesis next to each method. For each method, we use 15 different bins with an equal number of samples to group the error and the spread.
fect” calibration for almost every variable and lead time. The only exception is the air surface temperature, where EnAFNO even surpasses the ENS in terms of both forecast skill —up to a lead time of 3 days,— and calibration.

The spread-skill plots illustrated in Figure 5 are averaged globally making it difficult to assess the advantages and shortcomings of the different ensemble strategies per region. To address this, Figure 6 displays the error (solid) and spread (dashed) as a function of the latitude for (in columns) different lead times: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). The color lines represent the proposed ensemble (red), the IFS (blue), and the two benchmark probabilistic NWM (gray and green). We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in parenthesis next to each method.

Figure 6. Error (solid) and spread (dashed) of the different ensemble strategies as a function of the latitude. Displayed by rows are the surface zonal wind velocity, the air surface temperature, the geopotential height at 500 hPa, and the total column water vapour, whereas different lead times are displayed by columns: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). The color lines represent the proposed ensemble (red), the IFS (blue), and the two benchmark probabilistic NWM (gray and green). We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in parenthesis next to each method.
et al., 2020; Pathak et al., 2022) to compute the model error (see Appendix C4) and the spread (see Appendix C5). As in Figure 5, we observe an increase in both error and spread as a function of lead time regardless of the variable and ensemble strategy. However, the growth of these patterns varies across latitudes. Also, these latitudinal structures of the error/spread are similar across models and variables. For instance, for the surface zonal wind velocity at 1-day of lead time (Figure 6a), we observe small differences between models and latitudes, with error and spread values ranging from 0.5 to 1.5 m/s. At longer lead times we observe larger increments in the higher latitudes of both the Northern and Southern hemispheres than in the tropics. This structure is probably due to the (high-) low-variability of the atmosphere in the (high-latitudes) tropics. The results are promising since (data-driven) probabilistic NWMs mimic the error patterns of (physically-driven) NWP. However, we observe differences in the magnitude of the error between the ENS and the probabilistic NWMs, especially in the high latitudes, that are amplified with lead time, e.g., geopotential at 500 hPa panels in Figure 6 at lead times of 3-day and 5-day. These differences point to larger error structures in the NWMs ensembles than in ENS, and this is especially true for the G-AFNO ensemble. Finally, the EnAFNO latitudinal pattern of the spread closely approximates the error pattern, and more importantly, improves upon the G-AFNO that is systematically underdispersive across latitudes. Similar behaviour is reflected in the geopotential height at 500 hPa and the total column water vapour. However for the air surface temperature, we observe that EnAFNO even improves upon the benchmark IFS in terms of model error, shown previously in Figure 5, regardless of the latitude for 1-day and 3-days of lead time.

Figure 7 shows the CRPS spatial fields for the IFS (first column) and EnAFNO (second column) for the 4 variables considered (rows). The CRPS allows us to examine the overall probabilistic skill of the models (see Appendix C6), —the lower CRPS the better skill,— thus complementing the analysis presented in Figures 5 and 6. Also, we show the CRPSS (third column), which is a score that compares the CRPS between ensembles and has been computed relative to the IFS (see Appendix C7). In the right column, red (blue) colors define the areas where the CRPSS is above (below) 0. This allows us to identify the regions where EnAFNO has lower (higher) CRPS than the IFS, and therefore better (worse) probabilistic skill. For the CRPS we observe similar spatial patterns between the IFS and EnAFNO, which are very dependent on the dynamics of each variable. For instance, for the air surface temperature we find higher values over land than in the ocean, which is a direct consequence of the high (low) variability of this variable at land (the ocean). Other example is the total column water vapour, where the CRPS is high close to the Equator, the tropics and over the sea —which are areas with large amounts of moisture in the air,— and decreases with latitude and over land. According to the CRPSS, EnAFNO has overall better (worse) probabilistic skill than the IFS for the surface zonal wind velocity and the air surface temperature over land (the ocean), slightly worse for the total column water vapour, and worse for the geopotential at 500 hPa —except at the tropics. Therefore, the majority of these results are consistent with Figure 5 and Figure 6 for 3-day lead time, where overall IFS presented lower error and a better spread-skill relationship than EnAFNO when aggregated globally or by latitude, respectively. Nevertheless, by displaying the CRPS over 2D spatial fields in Figure 7, we show that EnAFNO has better probabilistic skill scores than IFS over land for key variables such as air surface temperature and surface wind, which are of direct interest to many sectors.

4.2 Evaluation of aleatoric uncertainty: breeding of growing modes

EnAFNO relies on bred vectors for the aleatoric uncertainty. As explained in Section 3.1.2, the BGM method aims to estimate the “first” guess errors which ultimately are representations of the growing components of the error field. While the growing components start to amplify from the beginning of the forecast, the nongrowing ones usually decay and they start to amplify only at a later time when they finally project onto
Figure 7. CRPS spatial fields for the IFS and EnAFNO (columns) and for the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa and the total column water vapour (rows) for a forecast lead time of 3 days. The lower the CRPS the better the probabilistic skill of the ensemble. Also shows is the CRPSS (third column). Red (blue) colors define areas where the EnAFNO has lower (higher) CRPS than IFS and therefore better (worse) probabilistic skill. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. See Appendix C6 and Appendix C7 for the details on the computation of the CRPS and CRPSS, respectively.

According to the explanation above, the spread of forecasts based on bred (Gaussian) perturbations should increase (decrease) with lead time. To test the latter assumption in the context of NWMs, we show in Figure 8 the spatial ensemble spread (hereafter just spread, see Appendix C5) for the surface zonal wind velocity (Figure 8.a), the air surface temperature (Figure 8.b), the geopotential at 500 hPa (Figure 8.c) and the total column water vapour (Figure 8.d) as a function of lead time for the G-AFNO (gray) and EnAFNO (blue) ensembles. We observe two different behaviours. At the most early lead times of the forecast (e.g., ≤ 24 hours in Figure 8) the spread of EnAFNO (G-AFNO) increases (decreases). This indicates that the bred (white) perturbations of EnAFNO (G-AFNO) are representations of the growing (non-growing) directions of the error pattern. At longer lead times of the forecast (e.g., 24 hours in Figure 8) the spread of both ensembles increases with time. While this seems natural for EnAFNO, for G-AFNO this
is because the error patterns have finally projected onto the growing modes after a few iterations.

**Figure 8.** Comparison between bred and gaussian noise in NWM predictive systems. Evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 2 ensembles: the G-AFNO (gray) and EnAFNO (blue) ensembles, to compare white and bred perturbations. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in brackets next to each method.

**4.3 Evaluation of epistemic uncertainty: model retraining and model sampling**

Figure 9 shows the forecast error (solid) and the spatial ensemble spread (dashed) for the surface zonal wind velocity (Figure 9.a), the air surface temperature (Figure 9.b), the geopotential at 500 hPa (Figure 9.c) and the total column water vapour (Figure 9.d) as a function of lead time. We represent two ensembles: AFNO-0 (pink) and AFNO-A (purple). Both AFNO-0 and AFNO-A are 30-member ensembles of forecasts, that have been generated by means of model sampling (MS). In particular, AFNO-0 follows the benchmark MS strategy, where models were collected right after model convergence by allowing the training to continue for a few epochs with a high learning rate (see the pink illustrations in Figure 3). Meanwhile, AFNO-A is based on the proposed MS strategy, where models are sampled after changing the loss surface, in addition to the computa-
tion of SGD with a high learning rate (see the purple illustrations in Figure 3). Therefore, Figure 9 allows for a comparison between model sampling strategies, and their ability to produce a diverse set of plausible forecasts. To isolate this aspect of EnAFNO, we fed the ensemble of NWMs with the non-perturbed initial conditions (i.e., standardized ERA5 fields).

![Comparison of model sampling strategies](image)

**Figure 9.** Comparison of model sampling strategies. Evolution of the spatial ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. Two ensembles are shown: AFNO-0 (pink) and AFNO-A (purple), to compare the benchmark and proposed epistemic uncertainty strategies, respectively. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in brackets next to each method.

As per the error, both AFNO-0 and AFNO-A show values that increase with lead time. Both methods produce similar error patterns up to day 3, where AFNO-A starts to increase the error at a lower rate than AFNO-0. Differently, the spread of AFNO-A grows with lead time while AFNO-0 increases at a very low rate. This enforces the model sampling strategy proposed to produce a diverse number of ensembles. Note that when AFNO-A is combined with the other uncertainty quantification strategies present in EnAFNO —model retraining and bred vectors,— the error (spread) decreases (increases) improving the calibration of the ensembles as shown in Figure 5.
5 Discussion & Conclusions

We proposed an ENsemble generation strategy based on bred vectors and Adaptive Fourier Neural Operators (EnAFNO), to build global ensembles of weather forecasts—although other Neural Weather Models (NWM) could be also used in practice. This strategy combines both aleatoric (initial condition error) and epistemic (model error) sources of uncertainty. For the aleatoric portion, we examined for the first time the Breeding of Growing Modes (BGM) method in NWM predictive systems, a technique traditionally used in several operational Numerical Weather Prediction (NWP) ones to estimate the initial condition uncertainty. For the epistemic uncertainty, we explored an alternative way to generate ensembles of NWMs for cases when model retraining is constrained by computational resources. We compared the resulting ensemble with the NWP from the European Centre for Medium-Range Weather Forecasts (ECMWF) that builds on the Integrated Forecast System (IFS), the gold standard in weather forecasting, and a benchmark probabilistic NWM based on white noise and the AFNO deep learning topology (G-AFNO). These two benchmarks are representative of the state-of-the-art in the NWP and NWM fields, respectively.

The design of EnAFNO implied optimizing a set of hyperparameters related to the proposed epistemic and aleatoric strategies. Figure A1 shows a set of binned spread-skill plots that represent different sensitivity tests to the epistemic and aleatoric uncertainties hyperparameters: the influence of the number of NWMs sampled (sampling epistemic uncertainty), the number of times we retrain the NWM from scratch (sampling epistemic uncertainty), the number of bred vectors and the amplitude of the initial perturbation in the breeding cycle (sampling aleatoric uncertainty). This optimization process led to an optimum 540-member ensemble based on the combination of 90 NWMs and 6 bred vectors per NWM. Note that including more members in the ensemble—either coming from a larger set of NWMs or bred perturbations,— than the ones indicated has no added value to the calibration skill, but can be potentially useful for other applications such as the exploration of extreme events. We require 30 seconds to produce a single 7-day global weather forecast, which is orders of magnitude faster than the time demanded by ECMWF’s IFS predictive system. With a cluster of 64 P100 GPUs available, we are capable of generating a ∼500-member ensemble in ∼5 minutes, which is 10 times more members than ECMWF’s IFS ensemble.

EnAFNO achieves lower error and higher spread than the NWM probabilistic benchmark G-AFNO, resulting in better-calibrated ensembles and approaching the calibration of the IFS. Moreover, the spatial representation of the errors is qualitatively similar to the IFS. This is encouraging since data-driven models, which are optimized statistical functions to the problem of weather forecasting, can mimic the error pattern of NWP tools, which are based on equations describing the dynamics of the atmosphere. Furthermore, the probabilistic skill of EnAFNO as given by the CRPS is better than the IFS over land for the air surface temperature and surface wind velocity. This extremely relevant for many socio-economic sectors (e.g., energy) that strongly rely on these variables. The success of EnAFNO builds on two pillars: the combination of model sampling and model retraining (epistemic uncertainty) and the BGM method (aleatoric uncertainty).

We proved that the BGM method is more effective than benchmark white-noise perturbations to increase the diversity of the ensemble (see Figure 8). This improvement is related to the ability of bred vectors to perturb along the unstable modes of the system, while arbitrary random perturbations project onto decaying modes, such as gravity waves (Lacarra & Talagrand, 1988). However, it is still to be explored whether these breeding modes are actually estimates of the initial condition uncertainty or if they include also uncertainties stemming from the NWM model. This is because the nature of the growing errors within ERA5 comes from “first guess” forecasts, which are ultimately computed with an NWP system. Thus, when driven with NWM the breeding cycle might not be an approximation to the analysis cycle. Furthermore, NWM are trained using (re-
analysis data and they might have assimilated some of the analysis errors in the statistical function. Therefore, bred vectors estimated with NWM might portray a more complex picture of the uncertainties present in the system than just a representation of the growing errors within the analysis. In this regard, “toy” experiments in pseudo-atmospheres and synthetic datasets may help to disentangle the uncertainties quantified in NWM-bred vectors. In addition, we found a lack of diversity across the bred vectors computed. This is characteristic of this type of perturbation since they all tend to approximate the leading Lyapunov vectors. However, there exist several modifications to the BGM standard method that often help diversify the ensemble of perturbations and would be worth testing them in future studies (Annan, 2004; Primo et al., 2008; Pazó et al., 2011). Also, there are other strategies aiming at capturing the initial condition uncertainty that might be worthy studying in the future in the context of NWM, such as singular vectors or perturbed observation techniques (Hamill et al., 2000). One alternative to generate the ensembles is to use the 10 ERA5 members derived from the EDA system (Isaksen et al., 2010) or lagged fields (Isaksen et al., 2010).

For the epistemic uncertainty, Scher & Messori (2021) showed that model retraining was an effective strategy for use-cases building on CNNs and low-resolution data, where the retraining of NWM demanded short times. This might not necessarily be the case when training the millions of parameters of novel NWMs. To alleviate the computational costs, in this study we proposed a novel model sampling strategy. We do this by retaining the models at the different epochs after changing the loss surface. We proved that this approach generates a more diverse ensemble than by sampling models after model convergence. Here we took the benefits of the 2-step FourCastNet’s training scheme to put to the test the proposed model sampling strategy. Nevertheless, this can be easily adapted to other state-of-the-art NWMs, such as GraphCast (Lam et al., 2023) or Pangu-Weather (Bi et al., 2023), by allowing the model to train several epochs after model convergence with a different loss function —for instance, one that minimizes the error also at a different lead time such as the one used during the fine-tuning phase. Finally, besides model retraining and model sampling, there is another relevant source of epistemic uncertainty not quantified in this study. This is related to the NWM deep learning architecture of choice. In this regard, multi-model ensembles of forecasts are yet to be explored in NWM weather forecasting, and present promising avenues of future research.

Finally, statistical post-processing can be applied to NWM-based weather forecasts to improve the calibration and forecast skill of the ensemble, analogous to their implementation upon NWP-based products (Hu et al., 2023). Also, post-processing tools can be used to derive variables or indices that are not included in the setting of the NWM model, such as wildfire risk or precipitation (Duncan et al., 2022).

Appendix A Optimization of EnAFNO

Figure A1 shows the spread-skill plots (see Appendix C3) of the surface zonal wind velocity (first row) and the geopotential at 500 hPa (second row) for a set of ensembles that aim to explore different possible configurations or parameters for the aleatoric and epistemic uncertainty strategies proposed. Note that we have analyzed the results for more variables than the ones appearing in Figure A1 with no remarkable differences in the conclusions, and therefore we only show these two for simplicity.

For the epistemic uncertainty, we evaluate the influence of the number of models sampled (first column) and the number of times we may train the NWMs from scratch (second column) in the spread-error relationship. To this aim, we include in Figures A1.a and A1.e 3 sets of ensembles with 5 (orange), 10 (red) and 30 (brown) members each. These figures show that model sampling can increase (decrease) the spread (error) of the ensemble but we have seen no major variations in these metrics beyond 30 NWMs. Similarly, in Figures A1.b and A1.f, we display 3 sets of ensembles which are the result of
stacking newly trained branches of sets of 30 NWMs: AFNO-A (turquoise), AFNO-A+B (blue), AFNO-A+B+C (dark blue). This allow us to examine the influence of model re-training on the spread and error values. We refer the reader to Section 3.1.4 for more information on the details of each AFNO, but basically the main difference among them is the training samples —1979-2015 for AFNO-A and AFNO-B, and 1979-2015 and 2019-2022 for AFNO-C. Figures A1.b and A1.f show that by training from scratch a new branch of 30 NWMs in identical conditions —i.e., same training/validation samples,— but with different random seeds for the initialization of the weight coefficients (AFNO-A and AFNO-B), we clearly increase (decrease) the spread (error), pushing the ensemble towards the ideal 1:1 ratio (diagonal line). In addition, still an even further but slight improvement on both metrics can still be achieved if we train a new branch in non-identical conditions —i.e., different training/validation samples (AFNO-C),— however it is not as remarkable as the previous one. This suggests that adding more branches of NWMs beyond a total of 3 —i.e., 90 NWMs in total,— has little to none added value to the spread-skill balance of the ensemble, even if we use a different set of training/validation samples.

As per the aleatoric, we evaluate the influence of the number and sign of bred vectors (third column), and the scale factor $k$ that determines the amplitude of the white noise during the first step of the breeding cycle (fourth column). We include in Figures A1.c and A1.g, 4 different ensembles: 1 bred vector with positive component (1+), 1 bred vector with both positive and negative components (1-), 3 bred vectors with positive component (3+), and 3 bred vectors with both positive and negative components (3+-). These 4 ensembles allow us to examine the change in spread and error if we increase the number of bred vectors (change the sign of the bred perturbation) by comparing the “1+” and “3+” (“1+” and “1+-”) curves. We discuss in Section 5 possible explanations to the lack of variability between bred vectors and potential avenues to increase their diversity.

Figure A1. Spread-skill plots of the surface zonal wind velocity (first row) and the geopotential at 500 hPa (second row) for a set of ensembles that aim to explore different aspects of EnAFNO aleatoric and epistemic uncertainty strategies. Next to each ensemble we put in brackets the number of members.

According to these results, we decided to (1) sample 30 models per branch, (2) train 3 branches in total, (3) use 3 bred vectors with both positive and negative components,
and (4) start the breeding cycle by sampling from a Gaussian distribution with 0 mean and 0.15 of standard deviation. The result is a 540-member ensemble which we named EnAFNO.

Appendix B Optimization of G-AFNO

Figure B1 shows the evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 5 versions of G-AFNO based on different number of members within the ensemble: 5 (red), blue (10), green (50), orange (100), gray (150). This figure allow us to examine the change in spread of the ensemble as a function of the number of members —or samples from the Gaussian perturbation distribution of 0 mean and 0.3 of standard deviation. We observe that the spread increases with lead time regardless of the number of members, except for the early lead times (see Section 4.2 for more details on this aspect). We also notice that the spread increases with the number of members. However, beyond a 50-member ensemble we do not report significant increments in the spread. This indicates that by perturbing the initial condition fields with 50 samples from a Gaussian distribution, we already achieve the maximum spread that the ensemble can achieve with this method.

Appendix C Definitions & Metrics

C1 Ensemble mean

We define the ensemble mean (EM) —at a given latitude, \( j = 1, 2, ..., N_{\text{lat}}, \) longitude, \( k = 1, 2, ..., N_{\text{lon}}, \) and lead time \( l = 1, 2, ..., L \)— as the average over the set of forecasts \( f_m \) initialized at a given date \( i, \) where \( m = 1, 2, ..., M \) is the number of members:

\[
EM_{i,j,k,l} = \frac{1}{M} \sum_{m} f_m
\]  
(C1)

C2 Ensemble spread

We define the ensemble spread (ES) —at a given latitude, \( j = 1, 2, ..., N_{\text{lat}}, \) longitude, \( k = 1, 2, ..., N_{\text{lon}}, \) and lead time \( l = 1, 2, ..., L \)— as the standard deviation of the set of forecasts \( f_m \) initialized at a given date \( i, \) where \( m = 1, 2, ..., M \) is the number of members:

\[
ES_{i,j,k,l} = \sqrt{\frac{1}{M} \sum_{m} (f_m - EM_{i,j,k,l})^2}
\]  
(C2)

C3 Binned spread-skill plots

Binned spread-skill plots are diagrams where the binned forecast skill is plotted against the binned ensemble spread. These plots allow to easily examine the calibration of the ensembles and have been widely used in probabilistic forecasting (Murphy, 1988; Leutbecher & Palmer, 2008; Hu et al., 2023). Therefore, if the spread-skill ratio is (close to) far away from the scale 1:1, then the ensemble is said to be (well-) ill-calibrated. We compute the binned skill following the next steps:

1. Compute the ensemble mean at every gridpoint, \( \{i, j\} \) forecast, \( i, \) and lead time, \( l, \) following Equation C1: \( EM_{i,j,k,l} \)
2. Concatenate the values in a single vector: \( \vec{EM} \)
3. Compute the squared of the differences between the forecasts, \( \vec{EM}, \) and ERA5, \( \vec{y}: \vec{E} = (\vec{EM} - \vec{y})^2 \)
4. Sort the differences \( \vec{E}. \)
5. Group the differences in equal-sized \( b = 1, 2, ..., B \) bins: \( \vec{E} = \{\vec{E}_1, \vec{E}_2, ..., \vec{E}_B\} \)
Figure B1. Evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 5 versions of G-AFNO based on different number of members within the ensemble: 5 (red), blue (10), green (50), orange (100), gray (150).

6. For every equal-sized bin of $s = 1, 2, ..., S$ samples, compute the mean such as:

$$E_1 = \frac{1}{S} \sum_{s} E_{1s}$$

7. Finally, compute the square root of the bin averaged values, $\vec{E} = \{E_1, E_2, ..., E_B\}$, such as:

$$\vec{E}^\frac{1}{2} = \{E_1^\frac{1}{2}, E_2^\frac{1}{2}, ..., E_B^\frac{1}{2}\}$$

For the computation of the binned spread we follow the ones below:

1. Compute the ensemble spread at every gridpoint, $\{i, j\}$ forecast, $i$, and lead time, $l$, following Equation C2: $ES_{i,j,k,l}$
2. Concatenate the values in a single vector: $ES$
3. Sort the spread vector, $ES$, based on the differences obtained during the computation of the binned skill in point 4: $ES'$.  
4. Group the differences in equal-sized $b = 1, 2, ..., B$ bins: $ES' = \{E_1', E_2', ..., E_B'\}$
5. For every equal-sized bin of $s = 1, 2, ..., S$ bins, compute the mean such as:

$$ES_1' = \frac{1}{S} \sum_{s} ES_{1s}'$$
Finally, compute the square root of the bin averaged values, $\overline{ES'} = \{ES'_1, ES'_2, ..., ES'_B\}$, such as: $\overline{ES'}^i = \{ES'^i_1, ES'^i_2, ..., ES'^i_B\}$.

C4 Root mean squared error

We choose the Root Mean Squared Error (RMSE) as one of our primary metrics to evaluate the methods since it has been widely utilized in previous studies (Rasp et al., 2020; Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023). The RMSE allow us to examine the error of the ensemble mean and thus the forecast skill. In particular, we follow the definition introduced in Rasp et al. (2020), a study that proposed an evaluation framework for data-driven weather models. Therefore, for a set of $i = 1, 2, ..., N$ forecasts:

$$\text{RMSE} = \frac{1}{N_f} \sum_i^{N_f} \sqrt{\frac{1}{N_{lat}N_{lon}} \sum_j^{N_{lat}} \sum_k^{N_{lon}} (f_{i,j,k} - y_{i,j,k})^2} \quad (C3)$$

Note that for Figure 6 we simply do not average across latitudes since the RMSE is computed for each one independently. This is also one key reason why we did not include the latitude-weighted factor appearing in Rasp et al. (2020) in Equation C3.

C5 Spatial ensemble spread

Different from the ensemble spread introduced in Section C2, which describes how to compute this metric for each gridpoint individually, the “spatial ensemble spread (SES)” is domain-wise. This means that the goal is to map the ensemble spread of a group of gridpoints to a single value. We imitate the expression introduced by Rasp et al. (2020) for the RMSE, and develop the following one for the SES for a set of $i = 1, 2, ..., N$ forecasts:

$$\text{SES} = \frac{1}{N_f} \sum_i^{N_f} \sqrt{\frac{1}{N_{lat}N_{lon}} \sum_j^{N_{lat}} \sum_k^{N_{lon}} (f_{i,j,k} - y_{i,j,k})^2} \quad (C4)$$

C6 Continuous Ranked Probabilistic Score (CRPS)

The Continuous Ranked Probabilistic Score is a metric that has been traditionally used to evaluate ensembles of forecasts (Hersbach, 2000; Matheson & Winkler, 1976). The CRPS establishes a comparison between the probability distribution as described by the ensemble members and the groundtruth (Hersbach, 2000; Matheson & Winkler, 1976). Therefore, the CRPS allows for an examination of the probabilistic aspect of the forecasts and thus, it perfectly complements any deterministic evaluation performed by means of the standard error (Section C4). If we consider a set of forecasts $f_m$ initialized at date $i = 1, 2, ..., N$, with $m = 1, 2, ..., M$ members, then the CRPS at a particular latitude, $j = 1, 2, ..., N_{lat}$, longitude, $k = 1, 2, ..., N_{lon}$, and lead time $l = 1, 2, ..., L$ is given by:

$$\text{CRPS}_{i,j,k,l} = \int_{-\infty}^{\infty} [F(f) - I(t \leq z)]^2 dz \quad (C5)$$

If we consider several forecasts initialized at dates $i = 1, 2, ..., N$, then the expression reduces to:

$$\text{CRPS}_{j,k,l} = \frac{1}{N} \sum_i^{N} \text{CRPS}_{i,j,k,l} \quad (C6)$$

We used the implementation in the CRPS Python package for the computation of the integral in Equation C5.
C7 Continuous Ranked Probabilistic Skill Score (CRPSS)

We follow Ghazvinian et al. (2021) and implement the Continuous Ranked Probability Skill Score (CRPSS) to assess the performance of a probabilistic forecast relative to a reference one:

\[
CRPSS_{j,k,l} = 1 - \frac{CRPS_{j,k,l}}{CRPS_{ref,j,k,l}} \tag{C7}
\]

Appendix D Open Research

ERA5 and IFS are open-source datasets readily accessible from the Copernicus climate data store (https://cds.climate.copernicus.eu/cdsapp#!/home) and TIGGE (https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=pf/) data archives, respectively. The codes for accessing data and the software developed in this study, are publicly accessible from the GitHub repository: https://github.com/CW3E/EnAFNO.

Acknowledgments

This work was supported by the California Department of Water Resources Atmospheric River Program Phase III (Grant 4600014294) and U.S. Army Corps of Engineers (US-ACE) Forecast Informed Reservoir Operations Award (USACE W912HZ1920023). Computations were performed using the COMET Supercomputer, which was made available by the Atmospheric River Program Phase 2 and 3 supported by the California Department of Water Resources (awards 4600013361 and 4600014294, respectively) and the Forecast Informed Reservoir Operations Program supported by the USACE Engineer Research and Development Center (award USACE W912HZ-15-2-0019). A.S. was partly supported by the National Aeronautics and Space Administration (Grant 80NSSC22K0926). We would like to acknowledge Dr. José Manuel Gutiérrez Llorente from the Santander Meteorology Group at the Institute of Physics of Cantabria for his helpful discussions about the breeding of growing modes method.

References


precipitation forecasts. Advances in Water Resources, 151, 103907.


Towards calibrated ensembles of neural weather model forecasts

J. Baño-Medina¹, A. Sengupta¹, D. Watson-Parris¹,², W. Hu³, L. Delle Monache¹⁴

¹Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA
²Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
³School of Integrated Sciences, James Madison University, Harrisonburg, Virginia

Key Points:
• We present a strategy to produce ensemble forecasts with Neural Weather Models (NWMs), accounting for aleatoric and epistemic uncertainties.
• The proposed strategy reports better probabilistic skill than benchmark NWM and is competitive with the Integrated Forecasting System (IFS).
• We generate a 540-member ensemble in 4-5 hours using 1 GPU and in just a few minutes with 64 GPUs—orders of magnitude faster than the IFS.

Corresponding author: Jorge Baño-Medina, jbanomedina@ucsd.edu
Abstract

Neural Weather Models (NWM) are novel data-driven weather forecasting tools based on neural networks that have recently achieved comparable deterministic forecast skill to current operational approaches using significantly less real-time computational resources. The short inference times required by NWMs allow the generation of a large ensemble potentially providing benefits in quantifying the forecast uncertainty, particularly for extreme events, which is of critical importance for various socio-economic sectors. Here we propose a novel ensemble design for NWMs spanning two main sources of uncertainty: epistemic —or model uncertainty,— and aleatoric —or initial condition uncertainty. For the epistemic uncertainty, we propose an effective strategy for creating a diverse ensemble of NWMs that captures uncertainty in key model parameters. For the aleatoric, we explore the “breeding of growing modes” for the first time on NWMs, a technique traditionally used for operational numerical weather predictions as an estimate of the initial condition uncertainty. The combination of these two types of uncertainty produces an ensemble of NWM-based forecasts that is shown to improve upon benchmark probabilistic NWM and is competitive with the 51-member ensemble of the European Centre for Medium-Range Weather Forecasts based on the Integrated Forecasting System (IFS) —a gold standard in weather forecasting,— in terms of both error and calibration. In addition, we report better probabilistic skill than the IFS over land for two key variables: surface wind and air surface temperature.

Plain Language Summary

Numerical weather predictions (NWP) are prone to different sources of errors that propagate through time and space and grow exponentially due to the chaotic nature of the atmosphere. These aspects have triggered the pursuit of probabilistic forecasting in an effort to produce an ensemble of weather forecasts that allows for a description of possible outcomes given an initial weather state. Current operational tools based on NWP models can provide a limited number of members within the ensemble, due to the large computational requirements, which may impact their ability to capture a priori all likely scenarios. Recently, Neural Weather Models (NWM) based on neural networks have been able to provide high-quality weather forecasts at a fraction of the real-time computational cost of NWP-based systems. However, the mechanisms to capture the uncertainty of the prediction of these systems and deriving ensembles of forecasts remain unexplored. Here we develop a strategy to generate a 540-member ensemble that improves upon benchmark probabilistic NWM and is competitive with the 51-member ensemble of the European Centre for Medium-Range Weather Forecasts based on the Integrated Forecasting System —a gold standard in weather forecasting.

1 Introduction

Reliable and accurate weather forecasts are continuously demanded by practitioners in different socio-economic sectors (e.g., agriculture, tourism, water management) for reliable decision planning, to identify potential hazards or to allocate resources effectively, among others. Given an initial condition, global weather deterministic forecasts can be produced by a model — dynamical, statistical, or a hybrid combination,— that emulates the evolution of the atmosphere. Traditionally, forecasts have been based on Numerical Weather Prediction (NWP) models. NWP solves numerically a set of differential equations and execute sub-grid parameterization schemes to produce 4D multivariate, global atmospheric forecasts (Bauer et al., 2015). They require large computational resources to generate a forecast.

Neural Weather Models (NWM) are novel data-driven tools based on neural networks that require only a fraction of the resources demanded by NWP to generate a real-time prediction, and therefore might be potential cost-effective alternatives to generate
fast high-quality weather forecasts. Despite the short inference times reported (see e.g., Pathak et al. (2022)), they do require significant off-line resources to be trained. This training consists in learning a function describing the relationship between two atmospheric states separated in time (e.g., 6-hours), typically using global reanalysis data, such as the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5; Hersbach et al. (2020)). Once trained, given an initial condition they produce a forecast one time-step ahead. To generate forecasts farther into the future, we can repeat the operation $N$ times using as initial condition the prediction of the previous time-step (auto-regressive mode). Recently several topologies based on either graph neural networks (Lam et al., 2023) or transformers (Bi et al., 2023; Pathak et al., 2022) have reported global deterministic forecast skill on par with benchmark operational NWP models, such as the ECMWF Integrated Forecast System (IFS), and improved upon previous configurations exclusively based on convolutional neural networks (Rasp et al., 2020; Rasp & Thuerey, 2021; Weyn et al., 2021; Scher & Messori, 2021).

Nevertheless, both NWP and NWM forecasts are prone to errors due to different sources of uncertainty in the system that can be broadly classified into two different categories: epistemic or model uncertainty, and aleatoric or initial condition uncertainty. Epistemic uncertainty arises from our lack of knowledge or deficiencies in the model formulation to represent the “true” evolution of the atmosphere. Aleatoric uncertainty evolves around errors in the initial condition due to, e.g., miss-calibration of atmospheric instrumentation, lack of data coverage, or approximation and assumptions in the data assimilation algorithms. In addition, the atmosphere is a chaotic system. This implies that even an infinitesimally small perturbation will grow in time, drastically altering the evolution of the atmosphere as compared to the non-perturbed trajectory (Lorenz, 1963; Selz & Craig, 2023). As a consequence, our ability to predict the future evolution of the atmosphere is limited in time.

These aspects have led to the development of probabilistic forecasting to generate ensemble of weather forecasts that allow for a description of possible outcomes given an initial weather state (Leutbecher & Palmer, 2008). This can be achieved by slightly perturbing either the initial condition (aleatoric uncertainty) or the model (epistemic uncertainty), where each member then represents a different plausible evolution of the weather system. Ideally, the spread of a well-calibrated ensemble (i.e., measured by computing the standard deviation across members) will match the pattern of the errors, which can be estimated with the root mean squared error (RMSE) of the ensemble mean. To achieve the latter is far from trivial and in the last decades different ways have been explored to properly inject uncertainty into the weather prediction system and derive well-calibrated ensemble forecasts. Regarding the aleatoric uncertainty, one key condition is that the perturbations fields should represent the error patterns present in the analysis in both magnitude and shape. Decades of NWP research have transitioned from white noise perturbations to either Singular Vectors (SV, Molteni & Palmer (1993); Mureau et al. (1993)) or Breeding of Growing Modes (BGM, Toth & Kalnay (1997)). In the area of NWM simple methods have been tested so far to perturb the initial condition fields (e.g., based on white noise (Pathak et al., 2022), perlin noise (Bi et al., 2023), pink noise (Grubner et al., 2022) or lagged fields (Brenowitz et al., 2024)). To incorporate model uncertainty in the ensemble forecast, NWP builds on a variety of techniques that include utilizing different parameterization schemes (Palmer et al., 2009), stochastic physics perturbation (Sanchez et al., 2016) or multi-model ensembles (Medina et al., 2018), among others. Ultimately, these techniques aim to represent different physics in the atmosphere.

In NWMs, model uncertainty can not be completely disentangled from the initial condition one, since models are ultimately trained using “imperfect” (re-)analysis data, and therefore they have also assimilated the latter source of error in the model coefficients. Some studies have performed model retraining to generate an ensemble of models. This methodology consists in starting from arbitrarily different initial configurations of coefficients in the optimization process (Scher & Messori, 2021). However, model retrain-
ing might be computationally demanding, since implies training from scratch every model
of the ensemble. This goes against the main reason behind the exploration of NWM for
weather forecasting, which is their computational benefits as opposed to NWP. A cost-
effective alternative to model retraining is model sampling, where ensembles of models
are built by selecting different sets of coefficients that correspond to different epochs in
the optimization process (Weyn et al., 2021; Graubner et al., 2022).

Therefore, ensemble forecasting has been explored extensively for NWP, but the
same is not true for NWM. There is a growing interest in the design of ensembles based
on NWMs due to their ability to produce real-time forecasts with much shorter times
than NWP, i.e., about 45,000 times faster than traditional NWP models on a node-hour
basis (Pathak et al., 2022). This speedup could significantly increase the number of en-
semble members generated and potentially revolutionize the field of probabilistic fore-
casting.

In this study, we propose an ensemble generation strategy based on both aleatoric
and epistemic sources of uncertainty. For the aleatoric uncertainty, we examine the BGM
approach for the first time on NWM. We select this method since it builds on current
practices in NWP modeling —as opposed the ones tested so far by the community,—
and can be easily implemented with NWM. Meanwhile, for the epistemic, we propose
an innovative strategy to build a diverse ensemble of NWM. We require 30 seconds to
produce a single 7-day global weather forecast, which is orders of magnitude faster than
the time demanded by ECMWF’s IFS predictive system. With a cluster of 64 P100 GPUs
available, we are capable of generating a ~500-member ensemble in ~5 minutes, which
is 10 times more members than ECMWF’s IFS ensemble.

2 Data

We use the ERA5 reanalysis to train the NWMs (Hersbach et al., 2020), which is
a state-of-the-art dataset provided by the European Center for Medium-Range Weather
Forecasts (ECMWF). Reanalysis datasets are combinations of observations and weather
re-forecasts through data assimilation, and nowadays provide the most accurate repre-
sentation of the atmosphere (Kalnay et al., 2018). The NWM models proposed so far
in the observational space have used this dataset to train the models (Rasp et al., 2020;
Rasp & Thuerey, 2021; Weyn et al., 2021; Scher & Messori, 2021; Pathak et al., 2022;
Bi et al., 2023; Lam et al., 2023). ERA5 provides hourly estimates for a large number
of ocean-wave, land-surface and atmospheric variables at different pressure levels at 0.25°
of spatial resolution, resulting into 780×1440 latitude-longitude fields. As a compari-
son to the NWMs, we also include in the analysis the 50-member ensemble from the ECMWF
that builds on the IFS (see Section 3.2 for more details).

3 Methods

3.1 EnAFNO: Ensemble based on bred vectors and adaptive fourier neu-
ral operators

Figure 1 highlights the proposed strategy to generate ensemble of forecasts. This
strategy builds on bred vectors for the aleatoric uncertainty and both model sampling
and model retraining strategies for the epistemic uncertainty. We use the Adaptive Fourier
Neural Operator (AFNO) deep learning topology or FourCastNet (Pathak et al., 2022)
to predict a set of atmospheric variables 6-hours into the future. We used this model since
the code was publicly available when we started working on this study and had mem-
ory requirements that fit in our computational infrastructure, but the strategy proposed
could also be used to generate ensembles of forecasts with other NWMs topologies, such
as Pangu-weather (Bi et al., 2023) or GraphCast (Lam et al., 2023). To produce fore-
casts at longer lead times, we perform inference in auto-regressive mode. The resulting
ensemble consists of 540 members in total, which is the result of perturbing the initial condition with 6 bred vectors and feed each of the perturbed fields to an ensemble of 90 AFNOs. We name this ENsemble generation strategy based on bred vectors and Adaptive Fourier Neural Operators: EnAFNO.

Figure 1.  Schematic of the proposed strategy to generate an ensemble forecast with NWMs. An initial condition field, defined by the state of a set of atmospheric variables at time \( t_0 \), is perturb with 6 bred vectors (aleatoric uncertainty) and then feed into 90 distinct NWMs (epistemic uncertainty), a 540-member ensemble. Note that for illustration purposes we only show one variable in the initial condition field, but it actually builds on multi-variable input fields.

In the following sections we will go through the details of AFNO (Section 3.1.1), the breeding of growing modes method (Section 3.1.2), and the (proposed) strategy to quantify epistemic uncertainty of NWMs (Section 3.1.3). Finally, in Section 3.1.4 we describe the training parameters and computational requirements of EnAFNO.

3.1.1 Neural weather model: Adaptive Fourier Neural Operator (AFNO)

FourCastNet is a deep neural network (see e.g., Goodfellow et al. (2016) for a review in deep learning) that combines Fourier Neural Operators (FNO, Li et al. (2020)) with a Vision Transformer (ViT, Li et al. (2020)) backbone. ViTs have revolutionized a number of computer vision applications thanks to the multi-head self-attention mechanism that allows the network to extract complex features from the training data. The Adaptive FNO (AFNO) was first introduced by Guibas et al. (2021) as a cost-effective solution to perform multi-head self-attention in high-resolution spaces, since the potential number of combinations within a ViT scales with the number of features, turning them impractical to train. AFNO tackles the latter by performing self-attention in the Fourier domain. The features now become frequencies, consequently reducing the number of possible combinations.

AFNO was further adopted by Pathak et al. (2022) for weather forecasting as FourCastNet, becoming the first global data-driven NWM to operate at 0.25° of spatial resolution, which is comparable to global operational NWP models. FourCastNet forecasts the next 6 hours of weather defined by 20 key atmospheric variables. To produce long rollouts of forecasts, FourCastNet can be run in auto-regressive mode. One key modification to the original AFNO that is of relevance to this manuscript, is the training scheme (see Figure 3). The original FourCastNet training consists of two steps: pre-training and
fine-tuning. During pre-training (fine-tuning) the network minimizes the error between the forecasted fields 6 hours (6 and 12 hours) ahead and the ground-truth. Therefore, fine-tuning prepares the network to operate in auto-regressive mode. The authors of the study report training and inference times of 16 hours and 2 seconds (for a week-long forecast) on a cluster of 64 A100 GPUs, respectively. Therefore, inference is 45,000 times faster than traditional NWP models on a node-hour basis and requires 12,000 times less energy. In terms of forecast skill, FourCastNet shows comparable results to the IFS. The skill falls down considerably beyond weather time-scales into the sub-seasonal regime. In this regard, new research outlines the potential of incorporating the Earth’s geometry into the NWM (Bonev et al., 2023). We refer the reader to the original manuscripts for more details on the model (see Guibas et al. (2021) and Pathak et al. (2022)).

3.1.2 Aleatoric uncertainty: breeding of growing modes

Breeding of growing modes (BGM) was a technique proposed by Toth & Kalnay (1997) to estimate initial perturbations that “reflect the analysis errors of the day” in the data. The errors typically appearing in modern (re-)analysis datasets, which are combinations of forecasts (“first guess”) and observations, can be classified into two different types: growing (e.g., baroclinically unstable modes) and non-growing (e.g., gravity waves). The “first guess” forecast favours the propagation of growing errors, which are maintained and keep breeding through the successive combinations of “first guess” forecasts and observations over time (analysis cycle). As a consequence, growing modes usually dominate the error pattern. BGM aims to estimate the uncertainty on the data by simulating how the growing errors are propagated in the analysis cycle. The breeding cycle consists of 6 different steps (see Figure 2): (1) initialize a random white noise perturbation and add it to the weather state at time \( t_0 = t \); (2) in parallel, feed the perturbed and (3) non-perturbed atmospheric fields to the NWM; (4) compute the differences between the perturbed and non-perturbed forecasts: (5) scale back the differences to the size of the initial perturbation; (6) repeat steps 1-to-5 but using the bred vector as a perturbation to the weather state at \( t_1 = t_0 + \Delta t \). The vector of (scaled) differences is usually referred to as the bred vector. Therefore, bred vectors are representations of the fastest growing modes of error, which are obtained by redistributing the initial white noise field. We can compute several bred vectors by initializing the cycle again with different random seeds. Note that the bred vectors are NWM-dependent, and thus we have to go through the breeding cycle for each NWM independently.

3.1.3 Epistemic uncertainty: model retraining and model sampling

The machine learning community has tested different strategies to estimate the epistemic uncertainty. For neural networks, the majority of these methods were traditionally based on Bayesian statistics, where prior distributions are placed upon the NN parameters (Neal, 2012). Then the posterior distribution is computed given the training data, ultimately providing an estimate of the predictive uncertainty. However, in modern neural networks we very often find millions of parameters, such as in the skillful NWM, and thus inferring the posterior distribution in these cases is computationally intractable. One solution is to approximate the posterior using dropout, a recurrent element in deep learning topologies that randomly activates/deactivates neurons within the network (Gal & Ghahramani, 2016). Another solution is model retraining, where ensembles of neural networks are trained by stochastically tuning different elements during training, e.g., initial model parameters, training samples. This approach has been shown to be very effective in many applications (Lakshminarayanan et al., 2017), including weather forecasting (Weyn et al., 2021; Scher & Messori, 2021). In particular, Scher & Messori (2021) proved that model retraining of CNNs produced the most diverse ensemble of forecasts as compared to Bayesian dropout, white noise and singular vectors. Nevertheless, this approach is extremely computationally demanding for modern NWM, since training a
single model may require up to 10 days depending on the computational infrastructure —i.e., number and type of GPUs. For this reason, a few studies have performed model sampling, which is based on the Stochastic Gradient Descent (SGD) iterations. SGD is the training algorithm used in neural networks to optimize the coefficients from an initial random state (Figure 3a) until convergence (Figure 3b), given a loss function. Therefore, the SDG iterations are repeated steps towards the direction of steepest descent in the loss surface, hence changing the initial model parameters. The size of the step is controlled by the learning rate, which is a tuneable parameter. An analysis of the model parameters at every epoch, suggests that the SGD trajectory contains useful information about the geometry of the posterior (Stephan et al., 2017; Maddox et al., 2019). Therefore, this methodology allows for efficient gathering of a large number of NWM by training only a single model from scratch and sampling models after convergence at different epochs (Weyn et al., 2021; Graubner et al., 2022). To maximize the diversity of the NWMs sampled, Stephan et al. (2017) suggested to place a high learning rate during this phase to enable a better exploration of the space of coefficients. One key drawback of this approach is that the models sampled not necessarily end up being diverse since SGD might simply oscillate towards a local/global minima (Figure 3d).

Here we propose a methodology to perform model sampling (Figure 3c), that builds on the 2-step training scheme that NWMs usually present: pre-training and fine-tuning. During pre-training (fine-tuning) the network minimizes the error between the forecasted fields in the next time step (two time steps) and the ground-truth. Therefore the loss function changes between training steps since the optimization objectives are different.
We propose to start sampling models after changing the loss function (at the beginning of the fine-tuning step). We hypothesize that this approach allows us to sample a diverse number of NWMs by exploring a large area of the space of solutions, since we are no longer trapped in a local/global minima as we are training in a new loss surface (Figure 3c).

We demonstrate this in Section 4.3.
3.1.4 Configuration of EnAFNO and computational requirements

We follow Pathak et al. (2022) and build on the following set of 20 variables from ERA5 at 00, 06, 12, 18 UTC: zonal and meridional wind velocities at 500, 850, 1000 hPa; geopotential height at 50, 500, 850, 1000 hPa; relative humidity at 500 and 850 hPa; air temperature at 500 and 850 hPa; surface temperature, mean sea level pressure and total column of water vapour. We scale the variables by subtracting the mean and dividing by the standard deviation over the training period. We follow the 2-step training procedure (i.e., pre-training and fine-tuning) described in Pathak et al. (2022). We use an Adam optimizer during the pre-training (fine-tuning) phase with a learning rate of 5E-4 and a cosine annealing scheduler (0.1 without scheduler) for 150 (35) epochs.

We inject epistemic uncertainty into the system by both model retraining and model sampling strategies. For model retraining, we train 3 branches of models, which we name AFNO-A, AFNO-B and AFNO-C. They are trained based on different seeds in the initialization of the model parameters and also different training periods: 1979-2015, 1979-2015 and both 1979-2015 and 2019-2022 for AFNO-A, AFNO-B and AFNO-C, respectively. For model sampling, we follow the strategy proposed in Section 3.1.3, and start sampling models during the fine-tuning step, beyond the epoch 156. Therefore, we leave a spin-up of 5 epochs to let the model adapt to the new error surface. In total we sample 30 models per branch since we observe no added value to the calibration skill beyond this quantity. We use the following nomenclature to refer to each of the models: AFNO$_{ij}$; where $i$ and $j$ represent the branch and number of sampled model within the branch, respectively (e.g., AFNO$_{A1}$).

To inject aleatoric uncertainty into the system we compute 3 independent breeding cycles per NWM (i.e., AFNO$_{ij}$). We initialize each of the cycles on the January 1st of 2018 with white noise of 0 mean and 0.15 of standard deviation. We let the cycle breed and maintain growing errors up to the 1st of March of 2018. For each breeding cycle we can then produce a pair of perturbations by either adding or subtracting the bred vector at time $t_0 = t$ to the corresponding (scaled) initial condition. Therefore, we can derive a total of 6 different members from the 3 independent breeding cycles per NWM.

Thus, the proposed ensemble of forecasts builds on 90 AFNO and 6 bred perturbations per NWM to derive a total of 540 members. This is almost 12 times more members than the IFS. Both the number of NWM models and bred vectors have been determined by testing different configurations (see Figure A1) and is further discussed in Sections 4.3 and 4.2. To train the models we used a cluster of 64 P100 GPUs. We report times of $\sim$ 7 days for training each branch, i.e., $\sim$ 21 days to derive the whole set of 90 AFNO, while to infer a 5-day 540-member global forecast on the same infrastructure takes $\sim$ 10 minutes. We note that shorter training/inference times than the ones presented here have been reported using better GPUs, such as the A100 (Pathak et al., 2022) or V100 (Bonev et al., 2023) type.

For illustration purposes, we show in Figure 4 the evolution of an atmospheric river over the Western coast of North America. We focus on this example, since it was already analyzed in the original FourCastNet study (Pathak et al., 2022) and is not included in the training data. Figure 4.a represents the total column water vapour for lead times of (from left to right) 6, 36, and 72 hours for (top to bottom) ERA5, and three random members of EnAFNO. We observe that EnAFNO is able to capture the main characteristics of the event showing consistent spatial patterns to those appearing in ERA5. Moreover, there are some differences among the three members shown. For instance, members 1 and 740 forecast larger total column water vapour values for the AR over the Pacific Ocean as compared to member 230 for a 36-hour of lead time. This exemplifies the variability of the NWM ensemble, that is able to represent different plausible evolutions of the atmospheric river. To visualize the performance of the entire ensemble, we average the to-
Figure 4. (a) Evolution of an atmospheric river over the Western coast of North America as represented by the total column water vapour. Model forecasts are initialized at 00:00 UTC on April 4, 2018. Example forecasts (from left to right) are shown for lead times 6, 36 and 72 hours and for ERA5 (row 1) and three random members of EnAFNO (rows 2-4); (b) Temporal series of the total column water vapour averaged over the Feather River Basin —latitudes \{39.0\textdegree, 41.5\textdegree\} and longitudes \{238.0\textdegree, 240.5\textdegree\},— for the 540 EnAFNO members (turquoise), the ensemble mean (blue) and ERA5 (black).

The total column water vapour over the Feather River basin —a water basin located north of the Sierras in California,— and display the evolution of the 7-day forecast as given by the members (turquoise), the ensemble mean of EnAFNO (blue) and ERA5 (black) in Figure 4.b. We observe how the ensemble mean is able to accurately reproduce the evolution of the atmospheric river up to 7-days of lead time. As expected, the ensemble spread increases as a function of lead time, and the groundtruth, i.e., ERA5, which is always within the range of possible EnAFNO evolutions.

---
3.2 Benchmark methods

We use the following benchmark models as comparison to EnAFNO:

• **IFS.** This ensemble builds on the ECMWF’s IFS (hereafter simply referred to as IFS). IFS consists of 50 members based on the Ensemble of Data Assimilation (EDA) system (Isaksen et al., 2010), that accounts for both model and initial condition uncertainty. It is considered to be the state-of-the-art NWP model in probabilistic forecasting. The forecasts are initialized every day at both 00:00 and 12:00 UTC, and have typically operated with spatial resolutions comparable to ERA5, what allows a direct comparison with the NWMs developed herein.

• **G-AFNO.** This ensemble was introduced in the original FourCastNet study as a first approximation to build ensemble of forecasts (Pathak et al., 2022). G-AFNO consists of 100 members, which are produced by perturbing the initial condition with 100 samples of a Gaussian distribution with 0 mean and 0.3 of standard deviation. These perturbed fields then feed a single AFNO, which has been trained following the guidelines described in Pathak et al. (2022). Therefore, this strategy only injects aleatoric uncertainty into the system. This ensemble represents the current benchmark involving the AFNO model. Since EnAFNO consists of 540 members, for a fair comparison we tested the sensitivity of the number of white-noise perturbations to the total spread of the ensemble, with no major differences beyond a 100-member ensemble (see Figure B1).

4 Results

We focus on four key atmospheric variables for different socio-economic sectors: the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa and the total column water vapour. We split the analysis into 3 different parts. First, we validate EnAFNO in terms of forecast skill and spread-error calibration (Section 4.1). Second, we evaluate the aleatoric uncertainty strategy of EnAFNO, i.e., the BGM method (Section 4.2). Third, we evaluate the proposed epistemic uncertainty strategy for EnAFNO and compare it against the approaches in prior literature (Section 4.3). We build on the following set of metrics (see Appendix C) to analyze the forecasts: standard error of the ensemble mean (skill), spatial ensemble spread (SES), binned spread-skill plots, and the Continuous Ranked Probabilistic Score (CRPS). We initialize the forecasts every day at 0:00 UTC from January 10th to the 28th of February of 2018. January 1st to January 10th are used to spin-up the bred vectors.

4.1 Evaluation of EnAFNO

We start by evaluating both the model error and model calibration of EnAFNO on a global scale. To this aim, Figure 5 displays the spread-skill plots for (in rows) the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa, and the total column water vapour at (in columns) different lead times: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). In colours we represent the 3 ensemble strategies: EnAFNO (blue), ENS (red), and G-AFNO (gray). The diagonal line represents the space of “perfect” calibration where the ensemble spread and the error are equal. If an ensemble has a higher (lower) error than spread this means that it is underdispersive (overdispersive), i.e., the probabilistic forecast is overconfident (underconfident). Please refer to Appendix C3 for details on the computation of these spread-skill plots. We observe that both the error and the spread of the forecasts increase as a function of lead time, regardless of the ensemble strategy considered. Nevertheless, the skill-spread ratio is different across variables, methods, and lead times. G-AFNO produces underdispersive ensembles, as already outlined in Graubner et al. (2022), and presents higher errors compared to other methods. Meanwhile, EnAFNO shows spread-skill val-
Figure 5. Binned spread-skill diagrams for the different ensemble strategies. Displayed by rows are the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa, and the total column water vapour, whereas the different lead times are displayed by columns: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). The colored lines represent the proposed ensemble EnAFNO (red), the IFS (blue) and the two benchmark probabilistic NWM (gray and green). We build ensemble forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in parenthesis next to each method. For each method, we use 15 different bins with an equal number of samples to group the error and the spread.

The colored lines that are closer to the 1:1 diagonal line than G-AFNO. This approximation to the state of “perfect” calibration is mostly due to the increase in spread, although we can observe also a decrease in the error as compared to G-AFNO. Nevertheless, the IFS is the ensemble showing the best spread-error relationships, approximating the state of “per-
fect” calibration for almost every variable and lead time. The only exception is the air surface temperature, where EnAFNO even surpasses the ENS in terms of both forecast skill —up to a lead time of 3 days,— and calibration.

Figure 6. Error (solid) and spread (dashed) of the different ensemble strategies as a function of the latitude. Displayed by rows are the surface zonal wind velocity, the air surface temperature, the geopotential height at 500 hPa, and the total column water vapour, whereas different lead times are displayed by columns: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). The color lines represent the proposed ensemble (red), the IFS (blue), and the two benchmark probabilistic NWM (gray and green). We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in parenthesis next to each method.

The spread-skill plots illustrated in Figure 5 are averaged globally making it difficult to assess the advantages and shortcomings of the different ensemble strategies per region. To address this, Figure 6 displays the error (solid) and spread (dashed) as a function of the latitude for (in columns) different lead times: 1-day (+12 to +24 hours), 3-day (+60 to +72 hours) and 5-day (+108 to +120 hours). We follow previous work (Rasp...
et al., 2020; Pathak et al., 2022) to compute the model error (see Appendix C4) and the spread (see Appendix C5). As in Figure 5, we observe an increase in both error and spread as a function of lead time regardless of the variable and ensemble strategy. However, the growth of these patterns varies across latitudes. Also, these latitudinal structures of the error/spread are similar across models and variables. For instance, for the surface zonal wind velocity at 1-day of lead time (Figure 6a), we observe small differences between models and latitudes, with error and spread values ranging from 0.5 to 1.5 m/s. At longer lead times we observe larger increments in the higher latitudes of both the Northern and Southern hemispheres than in the tropics. This structure is probably due to the (high-) low-variability of the atmosphere in the (high-latitudes) tropics. The results are promising since (data-driven) probabilistic NWMs mimic the error patterns of (physically-driven) NWP. However, we observe differences in the magnitude of the error between the ENS and the probabilistic NWMs, especially in the high latitudes, that are amplified with lead time, e.g., geopotential at 500 hPa panels in Figure 6 at lead times of 3-day and 5-day. These differences point to larger error structures in the NWMs ensembles than in ENS, and this is especially true for the G-AFNO ensemble. Finally, the EnAFNO latitudinal pattern of the spread closely approximates the error pattern, and more importantly, improves upon the G-AFNO that is systematically underdispersive across latitudes. Similar behaviour is reflected in the geopotential height at 500 hPa and the total column water vapour. However for the air surface temperature, we observe that EnAFNO even improves upon the benchmark IFS in terms of model error, shown previously in Figure 5, regardless of the latitude for 1-day and 3-days of lead time.

Figure 7 shows the CRPS spatial fields for the IFS (first column) and EnAFNO (second column) for the 4 variables considered (rows). The CRPS allows us to examine the overall probabilistic skill of the models (see Appendix C6), —the lower CRPS the better skill,— thus complementing the analysis presented in Figures 5 and 6. Also, we show the CRPSS (third column), which is a score that compares the CRPS between ensembles and has been computed relative to the IFS (see Appendix C7). In the right column, red (blue) colors define the areas where the CRPSS is above (below) 0. This allows us to identify the regions where EnAFNO has lower (higher) CRPS than the IFS, and therefore better (worse) probabilistic skill. For the CRPS we observe similar spatial patterns between the IFS and EnAFNO, which are very dependent on the dynamics of each variable. For instance, for the air surface temperature we find higher values over land than in the ocean, which is a direct consequence of the high (low) variability of this variable at land (the ocean). Other example is the total column water vapour, where the CRPS is high close to the Equator, the tropics and over the sea —which are areas with large amounts of moisture in the air,— and decreases with latitude and over land. According to the CRPSS, EnAFNO has overall better (worse) probabilistic skill than the IFS for the surface zonal wind velocity and the air surface temperature over land (the ocean), slightly worse for the total column water vapour, and worse for the geopotential at 500 hPa —except at the tropics. Therefore, the majority of these results are consistent with Figure 5 and Figure 6 for 3-day lead time, where overall IFS presented lower error and a better spread-skill relationship than EnAFNO when aggregated globally or by latitude, respectively. Nevertheless, by displaying the CRPS over 2D spatial fields in Figure 7, we show that EnAFNO has better probabilistic skill scores than IFS over land for key variables such as air surface temperature and surface wind, which are of direct interest to many sectors.

4.2 Evaluation of aleatoric uncertainty: breeding of growing modes

EnAFNO relies on bred vectors for the aleatoric uncertainty. As explained in Section 3.1.2, the BGM method aims to estimate the “first” guess errors which ultimately are representations of the growing components of the error field. While the growing components start to amplify from the beginning of the forecast, the nongrowing ones usually decay and they start to amplify only at a later time when they finally project onto
Figure 7. CRPS spatial fields for the IFS and EnAFNO (columns) and for the surface zonal wind velocity, the air surface temperature, the geopotential at 500 hPa and the total column water vapour (rows) for a forecast lead time of 3 days. The lower the CRPS the better the probabilistic skill of the ensemble. Also shows is the CRPSS (third column). Red (blue) colors define areas where the EnAFNO has lower (higher) CRPS than IFS and therefore better (worse) probabilistic skill. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. See Appendix C6 and Appendix C7 for the details on the computation of the CRPS and CRPSS, respectively.

According to the explanation above, the spread of forecasts based on bred (Gaussian) perturbations should increase (decrease) with lead time. To test the latter assumption in the context of NWMs, we show in Figure 8 the spatial ensemble spread (hereafter just spread, see Appendix C5) for the surface zonal wind velocity (Figure 8.a), the air surface temperature (Figure 8.b), the geopotential at 500 hPa (Figure 8.c) and the total column water vapour (Figure 8.d) as a function of lead time for the G-AFNO (gray) and EnAFNO (blue) ensembles. We observe two different behaviours. At the most early lead times of the forecast (e.g., ≤ 24 hours in Figure 8) the spread of EnAFNO (G-AFNO) increases (decreases). This indicates that the bred (white) perturbations of EnAFNO (G-AFNO) are representations of the growing (non-growing) directions of the error pattern. At longer lead times of the forecast (e.g., 24 hours in Figure 8) the spread of both ensembles increases with time. While this seems natural for EnAFNO, for G-AFNO this...
is because the error patterns have finally projected onto the growing modes after a few iterations.

**Figure 8.** Comparison between bred and Gaussian noise in NWM predictive systems. Evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 2 ensembles: the G-AFNO (gray) and EnAFNO (blue) ensembles, to compare white and bred perturbations. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in brackets next to each method.

### 4.3 Evaluation of epistemic uncertainty: model retraining and model sampling

Figure 9 shows the forecast error (solid) and the spatial ensemble spread (dashed) for the surface zonal wind velocity (Figure 9.a), the air surface temperature (Figure 9.b), the geopotential at 500 hPa (Figure 9.c) and the total column water vapour (Figure 9.d) as a function of lead time. We represent two ensembles: AFNO-0 (pink) and AFNO-A (purple). Both AFNO-0 and AFNO-A are 30-member ensembles of forecasts, that have been generated by means of model sampling (MS). In particular, AFNO-0 follows the benchmark MS strategy, where models were collected right after model convergence by allowing the training to continue for a few epochs with a high learning rate (see the pink illustrations in Figure 3). Meanwhile, AFNO-A is based on the proposed MS strategy, where models are sampled after changing the loss surface, in addition to the computa-
tion of SGD with a high learning rate (see the purple illustrations in Figure 3). Therefore, Figure 9 allows for a comparison between model sampling strategies, and their ability to produce a diverse set of plausible forecasts. To isolate this aspect of EnAFNO, we fed the ensemble of NWMs with the non-perturbed initial conditions (i.e., standardized ERA5 fields).

Figure 9. Comparison of model sampling strategies. Evolution of the spatial ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. Two ensembles are shown: AFNO-0 (pink) and AFNO-A (purple), to compare the benchmark and proposed epistemic uncertainty strategies, respectively. We build ensembles of forecasts for 50 initialization dates, from January 10th of 2018 at 00:00 UTC to February 28th of 2018. The size of the ensemble is indicated in brackets next to each method.

As per the error, both AFNO-0 and AFNO-A show values that increase with lead time. Both methods produce similar error patterns up to day 3, where AFNO-A starts to increase the error at a lower rate than AFNO-0. Differently, the spread of AFNO-A grows with lead time while AFNO-0 increases at a very low rate. This enforces the model sampling strategy proposed to produce a diverse number of ensembles. Note that when AFNO-A is combined with the other uncertainty quantification strategies present in EnAFNO—model retraining and bred vectors,—the error (spread) decreases (increases) improving the calibration of the ensembles as shown in Figure 5.
5 Discussion & Conclusions

We proposed an ENsemble generation strategy based on bred vectors and Adaptive Fourier Neural Operators (EnAFNO), to build global ensembles of weather forecasts —although other Neural Weather Models (NWM) could be also used in practice. This strategy combines both aleatoric (initial condition error) and epistemic (model error) sources of uncertainty. For the aleatoric portion, we examined for the first time the Breeding of Growing Modes (BGM) method in NWM predictive systems, a technique traditionally used in several operational Numerical Weather Prediction (NWP) ones to estimate the initial condition uncertainty. For the epistemic uncertainty, we explored an alternative way to generate ensembles of NWMs for cases when model retraining is constrained by computational resources. We compared the resulting ensemble with the NWP from the European Centre for Medium-Range Weather Forecasts (ECMWF) that builds on the Integrated Forecast System (IFS), the gold standard in weather forecasting, and a benchmark probabilistic NWM based on white noise and the AFNO deep learning topology (G-AFNO). These two benchmarks are representative of the state-of-the-art in the NWP and NWM fields, respectively.

The design of EnAFNO implied optimizing a set of hyperparameters related to the proposed epistemic and aleatoric strategies. Figure A1 shows a set of binned spread-skill plots that represent different sensitivity tests to the epistemic and aleatoric uncertainties: the influence of the number of NWMs sampled (sampling epistemic uncertainty), the number of times we retrain the NWM from scratch (sampling epistemic uncertainty), the number of bred vectors and the amplitude of the initial perturbation in the breeding cycle (sampling aleatoric uncertainty). This optimization process led to an optimum 540-member ensemble based on the combination of 90 NWMs and 6 bred vectors per NWM. Note that including more members in the ensemble —either coming from a larger set of NWMs or bred perturbations,— than the ones indicated has no added value to the calibration skill, but can be potentially useful for other applications such as the exploration of extreme events. We require 30 seconds to produce a single 7-day global weather forecast, which is orders of magnitude faster than the time demanded by ECMWF’s IFS predictive system. With a cluster of 64 P100 GPUs available, we are capable of generating a ∼500-member ensemble in ∼5 minutes, which is 10 times more members than ECMWF’s IFS ensemble.

EnAFNO achieves lower error and higher spread than the NWM probabilistic benchmark G-AFNO, resulting in better-calibrated ensembles and approaching the calibration of the IFS. Moreover, the spatial representation of the errors is qualitatively similar to the IFS. This is encouraging since data-driven models, which are optimized statistical functions to the problem of weather forecasting, can mimic the error pattern of NWP tools, which are based on equations describing the dynamics of the atmosphere. Furthermore, the probabilistic skill of EnAFNO as given by the CRPS is better than the IFS over land for the air surface temperature and surface wind velocity. This extremely relevant for many socio-economic sectors (e.g., energy) that strongly rely on these variables. The success of EnAFNO builds on two pillars: the combination of model sampling and model retraining (epistemic uncertainty) and the BGM method (aleatoric uncertainty).

We proved that the BGM method is more effective than benchmark white-noise perturbations to increase the diversity of the ensemble (see Figure 8). This improvement is related to the ability of bred vectors to perturb along the unstable modes of the system, while arbitrary random perturbations project onto decaying modes, such as gravity waves (Lacarra & Talagrand, 1988). However, it is still to be explored whether these breeding modes are actually estimates of the initial condition uncertainty or if they include also uncertainties stemming from the NWM model. This is because the nature of the growing errors within ERA5 comes from “first guess” forecasts, which are ultimately computed with an NWP system. Thus, when driven with NWM the breeding cycle might not be an approximation to the analysis cycle. Furthermore, NWM are trained using (re-
analysis data and they might have assimilated some of the analysis errors in the statistical function. Therefore, bred vectors estimated with NWM might portray a more complex picture of the uncertainties present in the system than just a representation of the growing errors within the analysis. In this regard, “toy” experiments in pseudo-atmospheres and synthetic datasets may help to disentangle the uncertainties quantified in NWM-bred vectors. In addition, we found a lack of diversity across the bred vectors computed. This is characteristic of this type of perturbation since they all tend to approximate the leading Lyapunov vectors. However, there exist several modifications to the BGM standard method that often help diversify the ensemble of perturbations and would be worth testing them in future studies (Annan, 2004; Primo et al., 2008; Pazó et al., 2011). Also, there are other strategies aiming at capturing the initial condition uncertainty that might be worthy studying in the future in the context of NWM, such as singular vectors or perturbed observation techniques (Hamill et al., 2000). One alternative to generate the ensembles is to use the 10 ERA5 members derived from the EDA system (Isaksen et al., 2010) or lagged fields (Isaksen et al., 2010).

For the epistemic uncertainty, Scher & Messori (2021) showed that model retraining was an effective strategy for use-cases building on CNNs and low-resolution data, where the retraining of NWM demanded short times. This might not necessarily be the case when training the millions of parameters of novel NWMs. To alleviate the computational costs, in this study we proposed a novel model sampling strategy. We do this by retaining the models at the different epochs after changing the loss surface. We proved that this approach generates a more diverse ensemble than by sampling models after model convergence. Here we took the benefits of the 2-step FourCastNet’s training scheme to put to the test the proposed model sampling strategy. Nevertheless, this can be easily adapted to other state-of-the-art NWMs, such as GraphCast (Lam et al., 2023) or PanguWeather (Bi et al., 2023), by allowing the model to train several epochs after model convergence with a different loss function — for instance, one that minimizes the error also at a different lead time such as the one used during the fine-tuning phase. Finally, besides model retraining and model sampling, there is another relevant source of epistemic uncertainty not quantified in this study. This is related to the NWM deep learning architecture of choice. In this regard, multi-model ensembles of forecasts are yet to be explored in NWM weather forecasting, and present promising avenues of future research.

Finally, statistical post-processing can be applied to NWM-based weather forecasts to improve the calibration and forecast skill of the ensemble, analogous to their implementation upon NWP-based products (Hu et al., 2023). Also, post-processing tools can be used to derive variables or indices that are not included in the setting of the NWM model, such as wildfire risk or precipitation (Duncan et al., 2022).

Appendix A Optimization of EnAFNO

Figure A1 shows the spread-skill plots (see Appendix C3) of the surface zonal wind velocity (first row) and the geopotential at 500 hPa (second row) for a set of ensembles that aim to explore different possible configurations or parameters for the aleatoric and epistemic uncertainty strategies proposed. Note that we have analyzed the results for more variables than the ones appearing in Figure A1 with no remarkable differences in the conclusions, and therefore we only show these two for simplicity.

For the epistemic uncertainty, we evaluate the influence of the number of models sampled (first column) and the number of times we may train the NWMs from scratch (second column) in the spread-error relationship. To this aim, we include in Figures A1.a and A1.e 3 sets of ensembles with 5 (orange), 10 (red) and 30 (brown) members each. These figures show that model sampling can increase (decrease) the spread (error) of the ensemble but we have seen no major variations in these metrics beyond 30 NWMs. Similarly, in Figures A1.b and A1.f, we display 3 sets of ensembles which are the result of
stacking newly trained branches of sets of 30 NWMs: AFNO-A (turquoise), AFNO-A+B (blue), AFNO-A+B+C (dark blue). This allow us to examine the influence of model retraining on the spread and error values. We refer the reader to Section 3.1.4 for more information on the details of each AFNO, but basically the main difference among them is the training samples —1979-2015 for AFNO-A and AFNO-B, and 1979-2015 and 2019-2022 for AFNO-C. Figures A1.b and A1.f show that by training from scratch a new branch of 30 NWMs in identical conditions —i.e., same training/validation samples,— but with different random seeds for the initialization of the weight coefficients (AFNO-A and AFNO-B), we clearly increase (decrease) the spread (error), pushing the ensemble towards the ideal 1:1 ratio (diagonal line). In addition, still an even further but slight improvement on both metrics can still be achieved if we train a new branch in non-identical conditions —i.e., different training/validation samples (AFNO-C),— however it is not as remarkable as the previous one. This suggests that adding more branches of NWMs beyond a total of 3 —i.e., 90 NWMs in total,— has little to none added value to the spread-skill balance of the ensemble, even if we use a different set of training/validation samples.

As per the aleatoric, we evaluate the influence of the number and sign of bred vectors (third column), and the scale factor $k$ that determines the amplitude of the white noise during the first step of the breeding cycle (fourth column). We include in Figures A1.c and A1.g, 4 different ensembles: 1 bred vector with positive component (1+), 1 bred vector with both positive and negative components (1-), 3 bred vectors with positive component (3+), and 3 bred vectors with both positive and negative components (3+-). These 4 ensembles allow us to examine the change in spread and error if we increase the number of bred vectors (change the sign of the bred perturbation) by comparing the “1+” and “3+” (“1+” and “1+-”) curves. We discuss in Section 5 possible explanations to the lack of variability between bred vectors and potential avenues to increase their diversity.

![Figure A1](image-url)  

**Figure A1.** Spread-skill plots of the surface zonal wind velocity (first row) and the geopotential at 500 hPa (second row) for a set of ensembles that aim to explore different aspects of EnAFNO aleatoric and epistemic uncertainty strategies. Next to each ensemble we put in brackets the number of members.

According to these results, we decided to (1) sample 30 models per branch, (2) train 3 branches in total, (3) use 3 bred vectors with both positive and negative components,
and (4) start the breeding cycle by sampling from a Gaussian distribution with 0 mean and 0.15 of standard deviation. The result is a 540-member ensemble which we named EnAFNO.

Appendix B Optimization of G-AFNO

Figure B1 shows the evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 5 versions of G-AFNO based on different number of members within the ensemble: 5 (red), blue (10), green (50), orange (100), gray (150). This figure allow us to examine the change in spread of the ensemble as a function of the number of members —or samples from the Gaussian perturbation distribution of 0 mean and 0.3 of standard deviation. We observe that the spread increases with lead time regardless of the number of members, except for the early lead times (see Section 4.2 for more details on this aspect). We also notice that the spread increases with the number of members. However, beyond a 50-member ensemble we do not report significant increments in the spread. This indicates that by perturbing the initial condition fields with 50 samples from a Gaussian distribution, we already achieve the maximum spread that the ensemble can achieve with this method.

Appendix C Definitions & Metrics

\textbf{C1 Ensemble mean}

We define the ensemble mean (EM) —at a given latitude, \( j = 1, 2, ..., N_{lat} \), longitude, \( k = 1, 2, ..., N_{lon} \), and lead time \( l = 1, 2, ..., L \) —as the average over the set of forecasts \( f_m \) initialized at a given date \( i \), where \( m = 1, 2, ..., M \) is the number of members:

\[ EM_{i,j,k,l} = \frac{1}{M} \sum_{m=1}^{M} (f_{i,j,k,l,m}) \]  \hfill (C1)

\textbf{C2 Ensemble spread}

We define the ensemble spread (ES) —at a given latitude, \( j = 1, 2, ..., N_{lat} \), longitude, \( k = 1, 2, ..., N_{lon} \), and lead time \( l = 1, 2, ..., L \) —as the standard deviation of the set of forecasts \( f_m \) initialized at a given date \( i \), where \( m = 1, 2, ..., M \) is the number of members:

\[ ES_{i,j,k,l} = \left( \frac{\sum_{m=1}^{M} (f_{i,j,k,l,m} - EM_{i,j,k,l})^2}{M} \right)^{1/2} \]  \hfill (C2)

\textbf{C3 Binned spread-skill plots}

Binned spread-skill plots are diagrams where the binned forecast skill is plotted against the binned ensemble spread. These plots allow to easily examine the calibration of the ensembles and have been widely used in probabilistic forecasting (Murphy, 1988; Leutbecher & Palmer, 2008; Hu et al., 2023). Therefore, if the spread-skill ratio is (close to) far away from the scale 1:1, then the ensemble is said to be (well-) ill-calibrated. We compute the binned skill following the next steps:

1. Compute the ensemble mean at every gridpoint, \( \{i,j\} \) forecast, \( i \), and lead time, \( l \), following Equation C1: \( EM_{i,j,k,l} \)
2. Concatenate the values in a single vector: \( \vec{EM} \)
3. Compute the squared of the differences between the forecasts, \( \vec{EM} \), and ERA5, \( \vec{y} \):
   \[ E = (\vec{EM} - \vec{y})^2 \]
4. Sort the differences \( \vec{E} \).
5. Group the differences in equal-sized \( b = 1, 2, ..., B \) bins: \( \vec{E} = \{\vec{E}_1, \vec{E}_2, ..., \vec{E}_B\} \)
Figure B1. Evolution of the ensemble spread as a function of forecast lead time for the (a) surface zonal wind velocity, (b) air surface temperature, (c) geopotential at 500 hPa, and (d) total column water vapour. We represent 5 versions of G-AFNO based on different number of members within the ensemble: 5 (red), blue (10), green (50), orange (100), gray (150).

6. For every equal-sized bin of \( s = 1, 2, \ldots, S \) samples, compute the mean such as:

\[
E_1 = \frac{1}{S} \sum_{s=1}^{S} E_{1s}
\]

7. Finally, compute the square root of the bin averaged values, \( \bar{E} = \{ E_1, E_2, \ldots, E_B \} \), such as:

\[
\bar{E} = \{ E_1^{\frac{1}{2}}, E_2^{\frac{1}{2}}, \ldots, E_B^{\frac{1}{2}} \}
\]

For the computation of the binned spread we follow the ones below:

1. Compute the ensemble spread at every gridpoint, \( \{i,j\} \) forecast, \( i \), and lead time, \( l \), following Equation C2: \( ES_{i,j,k,l} \)
2. Concatenate the values in a single vector: \( \bar{ES} \)
3. Sort the spread vector, \( \bar{ES} \), based on the differences obtained during the computation of the binned skill in point 4: \( \bar{ES}' \).
4. Group the differences in equal-sized \( b = 1, 2, \ldots, B \) bins: \( \bar{ES}' = \{ \bar{ES}'_1, \bar{ES}'_2, \ldots, \bar{ES}'_B \} \)
5. For every equal-sized bin of \( s = 1, 2, \ldots, S \) samples, compute the mean such as:

\[
\bar{ES}'_1 = \frac{1}{S} \sum_{s=1}^{S} \bar{ES}'_{1s}
\]
6. Finally, compute the square root of the bin averaged values, \( \overline{ES'} = \{ES'_1, ES'_2, ..., ES'_B\} \), such as: 
\[ \overline{ES'} = \{ES'_1, ES'_2, ..., ES'_B\} \]

C4 Root mean squared error

We choose the Root Mean Squared Error (RMSE) as one of our primary metrics to evaluate the methods since it has been widely utilized in previous studies (Rasp et al., 2020; Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023). The RMSE allow us to examine the error of the ensemble mean and thus the forecast skill. In particular, we follow the definition introduced in Rasp et al. (2020), a study that proposed an evaluation framework for data-driven weather models. Therefore, for a set of \( i = 1, 2, ..., N \) forecasts:

\[
RMSE = \frac{1}{N_f} \sum_i^{N_f} \sqrt{\frac{1}{N_{lat}N_{lon}} \sum_j^{N_{lat}} \sum_k^{N_{lon}} (f_{i,j,k} - y_{i,j,k})^2} \quad (C3)
\]

Note that for Figure 6 we simply do not average across latitudes since the RMSE is computed for each one independently. This is also one key reason why we did not include the latitude-weighted factor appearing in Rasp et al. (2020) in Equation C3.

C5 Spatial ensemble spread

Different from the ensemble spread introduced in Section C2, which describes how to compute this metric for each gridpoint individually, the “spatial ensemble spread (SES)” is domain-wise. This means that the goal is to map the ensemble spread of a group of gridpoints to a single value. We imitate the expression introduced by Rasp et al. (2020) for the RMSE, and develop the following one for the SES for a set of \( i = 1, 2, ..., N \) forecasts:

\[
SES = \frac{1}{N_f} \sum_i^{N_f} \sqrt{\frac{1}{N_{lat}N_{lon}} \sum_j^{N_{lat}} \sum_k^{N_{lon}} (f_{i,j,k} - y_{i,j,k})^2} \quad (C4)
\]

C6 Continuous Ranked Probabilistic Score (CRPS)

The Continuous Ranked Probabilistic Score is a metric that has been traditionally used to evaluate ensembles of forecasts (Hersbach, 2000; Matheson & Winkler, 1976). The CRPS establishes a comparison between the probability distribution as described by the ensemble members and the groundtruth (Hersbach, 2000; Matheson & Winkler, 1976). Therefore, the CRPS allows for an examination of the probabilistic aspect of the forecasts and thus, it perfectly complements any deterministic evaluation performed by means of the standard error (Section C4). If we consider a set of forecasts \( f_m \) initialized at date \( i = 1, 2, ..., N \), with \( m = 1, 2, ..., M \) members, then the CRPS at a particular latitude, \( j = 1, 2, ..., N_{lat} \), longitude, \( k = 1, 2, ..., N_{lon} \), and lead time \( l = 1, 2, ..., L \) is given by:

\[
CRPS_{i,j,k,l} = \int_{-\infty}^{\infty} [F(f) - I(t \leq z)]^2 dz \quad (C5)
\]

If we consider several forecasts initialized at dates \( i = 1, 2, ..., N \), then the expression reduces to:

\[
CRPS_{j,k,l} = \frac{1}{N} \sum_i^N CRPS_{i,j,k,l} \quad (C6)
\]

We used the implementation in the CRPS Python package for the computation of the integral in Equation C5.
C7 Continuous Ranked Probabilistic Skill Score (CRPSS)

We follow Ghazvinian et al. (2021) and implement the Continuous Ranked Probability Skill Score (CRPSS) to assess the performance of a probabilistic forecast relative to a reference one:

\[
CRPSS_{j,k,l} = 1 - \frac{CRPS_{j,k,l}}{CRPS_{j,k,l}^{ref}}
\]  

Appendix D Open Research

ERA5 and IFS are open-source datasets readily accessible from the Copernicus climate data store (https://cds.climate.copernicus.eu/cdsapp#!/home) and TIGGE (https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=pf/) data archives, respectively. The codes for accessing data and the software developed in this study, are publicly accessible from the GitHub repository: https://github.com/CW3E/EnAFNO.

Acknowledgments

This work was supported by the California Department of Water Resources Atmospheric River Program Phase III (Grant 4600014294) and U.S. Army Corps of Engineers (US-ACE) Forecast Informed Reservoir Operations Award (USACE W912HZ1920023). Computations were performed using the COMET Supercomputer, which was made available by the Atmospheric River Program Phase 2 and 3 supported by the California Department of Water Resources (awards 4600013361 and 4600014294, respectively) and the Forecast Informed Reservoir Operations Program supported by the USACE Engineer Research and Development Center (award USACE W912HZ-15-2-0019). A.S. was partly supported by the National Aeronautics and Space Administration (Grant 80NSSC22K0926). We would like to acknowledge Dr. José Manuel Gutiérrez Llorente from the Santander Meteorology Group at the Institute of Physics of Cantabria for his helpful discussions about the breeding of growing modes method.

References


precipitation forecasts. *Advances in Water Resources*, 151, 103907.


