`CoDiCast`: Conditional Diffusion Model for Global Weather Forecasting with Uncertainty Quantification

Jimeng Shi[1] , **Bowen Jin**[2] , **Jiawei Han**[2] , **Sundararaman Gopalakrishnan**[3] , **Giri Narasimhan**[1]

[1]Florida International University
[2]University of Illinois Urbana-Champaign
[3]National Oceanic and Atmospheric Administration
{jshi008, giri}@fiu.edu,
{bowenj4, hanj}@illinois.edu, sundararaman.g.gopalakrishnan@noaa.gov

Abstract

Accurate weather forecasting is critical for science and society. However, existing methods have not achieved the combination of high accuracy, low uncertainty, and high computational efficiency simultaneously. On one hand, traditional numerical weather prediction (NWP) models are computationally intensive because of their complexity. On the other hand, most machine learning-based weather prediction (MLWP) approaches offer efficiency and accuracy but remain deterministic, lacking the ability to capture forecast uncertainty. To tackle these challenges, we propose a conditional diffusion model, CoDiCast, to generate global weather predictions, integrating accuracy and uncertainty quantification at a modest computational cost. The key idea behind the prediction task is to generate realistic weather scenarios at a future time point, conditioned on observations from the recent past. Due to the probabilistic nature of diffusion models, they can be properly applied to capture the uncertainty of weather predictions. Therefore, we accomplish uncertainty quantifications by repeatedly sampling from stochastic Gaussian noise for each initial weather state and running the denoising process multiple times. Experimental results demonstrate that CoDiCast outperforms several existing MLWP methods in accuracy, and is faster than NWP models in inference speed. Our model can generate 6-day global weather forecasts, at 6-hour steps and 5.625-degree latitude-longitude resolutions, for over 5 variables, in about 12 minutes on a commodity A100 GPU machine with 80GB memory. The source code is available at https://github.com/JimengShi/CoDiCast.

1 Introduction

Weather prediction describes how the weather states evolve by mapping the current weather states to future weather states [Palmer, 2012]. Accurate weather forecasting is crucial for a wide range of societal activities, from daily planning to disaster preparedness [Shi *et al.*, 2025]. For example, governments, organizations, and individuals rely heavily on weather forecasts to make informed decisions that can significantly impact safety, economic efficiency, and overall well-being. However, weather predictions are intrinsically uncertain largely due to the complex and chaotic nature of atmospheric processes [Slingo and Palmer, 2011]. Therefore, assessing the range of probable weather scenarios is significant, enabling informed decision-making.

Traditional numerical weather prediction (NWP) methods achieve weather forecasting by approximately solving the differential equations representing the integrated system between the atmosphere, land, and ocean [Price *et al.*, May 2024; Nguyen *et al.*, 2023]. However, running such an NWP model can produce only one possibility of the forecast, which ignores the weather uncertainty. To solve this problem, *Ensemble forecast*[1] of multiple models is often employed to model the probability distribution of different future weather scenarios [Palmer, 2019; Leinonen *et al.*, 2023]. While such NWP-based ensemble forecasts effectively model the weather uncertainty, they have two primary limitations: physics-based models inherently make restrictive assumptions of atmospheric dynamics, and running multiple NWP models requires extreme computational costs [Rodwell and Palmer, 2007].

In recent years, machine learning (ML)-based weather predictions (MLWP) have been proposed to challenge NWP-based prediction methods [Ben Bouallègue *et al.*, 2024; Bülte *et al.*, 2024]. They have achieved enormous success with comparable accuracy and a much (usually three orders of magnitude) lower computational overhead. They are typically trained to learn weather patterns from a huge amount of historical data and predict the mean of the probable trajectories by minimizing the mean squared error (MSE) of model forecasts [Hewage *et al.*, 2021]. Representative work includes Pangu [Bi *et al.*, 2023], GraphCast [Lam *et al.*, 2023], ClimaX [Nguyen *et al.*, 2023], ForeCastNet [Pathak *et al.*, 2022], Fuxi [Chen *et al.*, 2023b], Fengwu and [Chen *et al.*, 2023a]. Despite the notable achievements of these MLWP methods, most of them are deterministic [Kochkov *et al.*, 2024], falling short in capturing the uncertainty in weather forecasts [Jaseena and Kovoor, 2022].

---

[1]Generating a set of forecasts, each of which represents a single possible scenario.

To compute the uncertainty for ML models, two methods exist. Perturbing initial conditions [Morley *et al.*, 2018] helps estimate the aleatoric uncertainty (data noise), while the Monte Carlo Dropout approach [Gal and Ghahramani, 2016] estimates epistemic uncertainty (model uncertainty) [Siddique *et al.*, 2022]. However, neither approach fully captures uncertainty in both the input conditions and the evolution of weather models. Additionally, these methods require manual tuning of perturbations and dropout rates, which can negatively impact model accuracy. These limitations motivate us to explore an approach for comprehensive uncertainty quantification without the loss of accuracy.

Denoising diffusion probabilistic models (DDPMs) [Ho *et al.*, 2020] stand out as a probabilistic type of generative model, which can generate high-quality images. By explicitly and iteratively modeling the noise additive and its removal, DDPMs can capture intricate details and textures of images. Furthermore, controllable diffusion models [Zhang *et al.*, 2023] enable the generation process to be guided by specific attributes or conditions, e.g., class labels, textual descriptions, or other auxiliary information. By doing so, the models can generate images that adhere to the specified conditions. This inspires us to consider the weather "prediction" tasks as "generation" tasks - generating plausible weather scenarios with conditional diffusion models. Promising potentials could be the following: (1) Weather numerical data is usually a 2-D grid over latitude and longitude, sharing a similar modality with the image. Diffusion models can capture the intricate weather distribution with iterative denoising. (2) Weather states from the recent past (i.e., initial conditions) can be injected into diffusion models to guide the generation of future weather evolution. (3) More notably, the starting noise sampling from the Gaussian distribution can mimic the aleatoric uncertainty while iteratively adding and removing noise captures the epistemic uncertainty. These features prompt probabilistic diffusion models to generate a set of diverse weather scenarios rather than a single deterministic one. Our contributions are presented as follows:

- We identify the shortcomings of current weather prediction methods. NWP-based methods are limited to restrictive assumptions and computationally intensive. Moreover, a single deterministic NWP- and MLWP-based method cannot achieve uncertainty quantification.

- To address these problems, we propose `CoDiCast`, a conditional diffusion model for global weather prediction conditioning on observations from the recent past while probabilistically modeling the uncertainty. In addition, we use the cross-attention mechanism to effectively integrate conditions into the denoising process to guide the generation tasks.

- We conduct extensive experiments on a decade of ERA5 reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECMWF), and evaluate our method against several state-of-the-art models in terms of accuracy, efficiency, and uncertainty. It turns out that `CoDiCast` achieves an essential trade-off among these valuable properties.

2 Related Work

Physics-based Numerical Weather Prediction. Numerical Weather Prediction (NWP) methods achieve weather forecasts by modeling the system of the atmosphere, land, and ocean with complex differential equations [Bauer *et al.*, 2015]. For example, High-Resolution Forecasts System (HRES) [ECMWF, 2023] forecasts possible weather evolution out to 10 days ahead. However, it is a deterministic NWP method that only provides a single forecast. To overcome the limitation of deterministic methods, the ensemble forecast suite (ENS) [Buizza, 2008] was developed as an ensemble of 51 forecasts by the European Centre for Medium-Range Weather Forecasts (ECMWF). ENS provides a range of possible future weather states, allowing for investigation of the detail and uncertainty in the forecast. Even if NWP ensemble forecasts effectively model the weather evolution, they exhibit sensitivity to structural discrepancies across models and high computational demands [Balaji *et al.*, 2022].

ML-Based Weather Prediction. Pangu [Bi *et al.*, 2023] employed three-dimensional transformer networks and Earth-specific priors to deal with complex patterns in weather data. GraphCast [Lam *et al.*, 2023] achieved medium-range weather prediction by utilizing an "encode-process-decode" configuration with each part implemented by graph neural networks (GNNs). GNNs perform effectively in capturing the complex relationship between a set of surface and atmospheric variables. A similar GNN-based work is [Keisler, 2022]. Fuxi [Chen *et al.*, 2023b] and Fengwu [Chen *et al.*, 2023a] also employ the "encode-decode" strategy but with the transformer-based backbone. FourCastNet [Pathak *et al.*, 2022] applied Vision Transformer (ViT) and Adaptive Fourier Neural Operators (AFNO), while ClimaX [Nguyen *et al.*, 2023] also uses a ViT backbone but the trained model can be fine-tuned to various downstream tasks. However, these models fall short in modeling the uncertainty of weather evolution [Jaseena and Kovoor, 2022] even though perturbations are added to initial conditions [Bülte *et al.*, 2024] and dropout methods [Gal and Ghahramani, 2016] are used to mimic the uncertainty. Additionally, ClimODE [Verma *et al.*, 2024] incorporated the physical knowledge and developed a continuous-time neural advection PDE weather model.

Diffusion Models. Diffusion models [Ho *et al.*, 2020] have shown their strong capability in computer vision tasks, including image generation [Li *et al.*, 2022], image editing [Nichol *et al.*, 2021], semantic segmentation [Brempong *et al.*, 2022] and point cloud completion [Luo and Hu, 2021]. Conditional diffusion models [Ho and Salimans, 2022] were later proposed to make the generation step conditioned on the current context or situation. However, not many efforts have adopted diffusion models in global medium-range weather forecasting. More recent research has focused on precipitation nowcasting [Asperti *et al.*, 2023; Gao *et al.*, 2024], and is localized in its predictions. GenCast [Price *et al.*, May 2024] is a recently proposed conditional diffusion-based ensemble forecasting for medium-range weather prediction. However, their conditioning is to directly use the observations from the recent past, which is shown to be insufficient (see the last case in the ablation study).

3 Preliminaries

In this section, we formulate the global weather prediction problem and briefly review Denoising Diffusion Probabilistic Models (DDPMs) [Ho *et al.*, 2020].

3.1 Problem Formulation

Deterministic Global Weather Predictions. Given the input consisting of the weather state(s), $X^t \in \mathbb{R}^{H \times W \times C}$ at time t, the problem is to predict a point-valued weather state, $X^{t+\Delta t} \in \mathbb{R}^{H \times W \times C}$ at a future time point $t + \Delta t$. $H \times W$ refers to the spatial resolution of data which depends on how densely we grid the globe over latitudes and longitudes, C refers to the number of channels (i.e., weather variables), and the superscripts t and $t + \Delta t$ refer to the current and future time points. The long-range multiple-step forecasts could be achieved by autoregressive modeling or direct predictions.

Probabilistic Global Weather Predictions. Unlike the deterministic models that output point-valued predictions, probabilistic methods model the probability of future weather state(s) as a distribution $P(X^{t+\Delta t} \mid X^t)$, conditioned on the state(s) from the recent past. Probabilistic predictions are appropriate for quantifying the forecast uncertainty and making informed decisions.
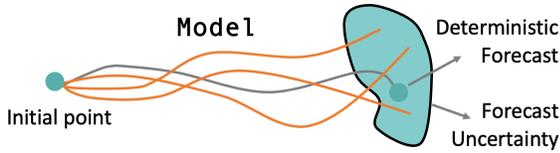

Figure 1: Deterministic vs Probabilistic Models.

3.2 Denoising Diffusion Probabilistic Models

A denoising diffusion probabilistic model (DDPM) [Ho *et al.*, 2020] generates target samples by learning a distribution $p_\theta(x_0)$ that approximates the target distribution $q(x_0)$. DDPM comprises a *forward diffusion* process and a *reverse denoising* process. The *forward process* transforms an input x_0 with a data distribution of $q(x_0)$ to a Gaussian noise vector x_N in N diffusion steps. It can be described as a Markov chain that gradually adds Gaussian noise to the input according to a variance schedule $\{\beta_1, \dots, \beta_N\}$:

$$q(x_n \mid x_{n-1}) = \mathcal{N}\left(x_n; \sqrt{1 - \beta_n}x_{n-1}, \beta_n \mathbf{I}\right), \text{and}$$

$$q(x_{1:N} \mid x_0) = \prod_{n=1}^{N} q(x_n \mid x_{n-1}), n \in [1, N],$$

where at each step n, the diffused sample x_n is obtained from x_{n-1} as described above. Multiple steps of the forward process can be described as follows in a closed form:

$$q(x_n \mid x_0) = \mathcal{N}\left(x_n; \sqrt{\bar{\alpha}_n}x_0, (1 - \bar{\alpha}_n)\mathbf{I}\right),$$

where $\alpha_n = 1 - \beta_n$ and $\bar{\alpha}_n = \prod_{s=1}^{n} \alpha_s$. Thus, $x_n = \sqrt{\bar{\alpha}_n}x_0 + \sqrt{1 - \bar{\alpha}_n}\epsilon$, with ϵ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

In the *reverse process*, the *denoiser* network is used to recover x_0 by stepwise denoising starting from the pure noise sample, x_N. This process is formally defined as:

$$p_\theta(x_{0:N}) = p(x_N) \prod_{n=1}^{N} p_\theta(x_{n-1} \mid x_n), \qquad (1)$$

where $p_\theta(x_n)$ is the distribution at step n parameterized by θ.

For each iteration, $n \in [1, N]$, diffusion models are trained to minimize the following KL-divergence:

$$\mathcal{L}_n = D_{KL}\left(q(x_{n-1} \mid x_n) \,||\, p_\theta(x_{n-1} \mid x_n)\right). \qquad (2)$$

where $q(x_{n-1}|x_n)$ can be computed from $q(x_n|x_{n-1})$ using Bayes rule and the multistep forward process equation above.

4 Methodology

This section introduces our approach for global weather prediction, `CoDiCast`, implemented as a conditional diffusion model. The key idea is to consider "prediction" tasks as "generation" tasks while conditioning on the context provided by recent past observation(s). An overview of the proposed `CoDiCast` is shown in Figure 2.

4.1 Forward Diffusion Process

At time point t, the forward diffusion process assumes that a pure noise sample X_N^{t+1} is generated from $X_0^{t+1} \in \mathbb{R}^{H \times W \times C}$ by adding noise N times (see the dotted lines in Figure 2):

$$X_n^{t+1} = \sqrt{\bar{\alpha}_n} \cdot X_0^{t+1} + \sqrt{1 - \bar{\alpha}_n}\epsilon, \qquad (3)$$

where ϵ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with the same dimensions as X_0^{t+1}, and $\bar{\alpha}$ is as described in Section 3.2.

4.2 Reverse Conditional Denoising Process

`CoDiCast` models the probability distribution of the future weather state, conditioned on the current and previous weather states. We exploit a pre-trained encoder to learn conditions as embedding representations of the past observations X^{t-1} and X^t, which are used to control and guide the synthesis process. Working in the latent space of embeddings works better than the original space of the observations.

$$p_\theta(X_{0:N}^{t+1} \mid \tilde{Z}^{t-1:t}) = p(X_N^{t+1}) \prod_{n=1}^{N} p_\theta(X_{n-1}^{t+1} \mid X_n^{t+1}, \tilde{Z}^{t-1:t}),$$

$$(4)$$

where $X_N^{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\tilde{Z}^{t-1:t}$ is the embedding representation as shown in Eq. (6).

After prediction at the first time point is obtained, a forecast trajectory, $X^{1:T}$, of length T, can be auto-regressively modeled by conditioning on the predicted "previous" states.

$$p_\theta(X_{0:N}^{1:T}) = \prod_{t=1}^{T} p(X_N^t) \prod_{n=1}^{N} p_\theta(X_{n-1}^t \mid X_n^t, \tilde{Z}^{t-2:t-1}). \quad (5)$$
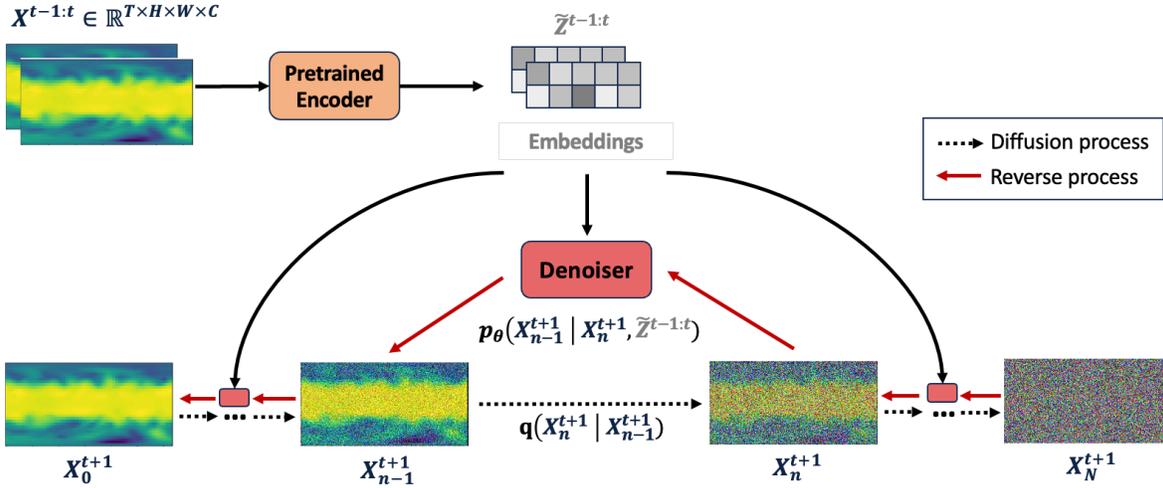
Figure 2: Framework of our conditional diffusion model for global weather forecast CoDiCast. The superscript T and the subscript N denote the time point and iteration step of adding/denoising noise. H and W represent the height (#latitude) and width (#longitude) of grid data. C is the number of variables of interest. X is the observation data and $\tilde{Z}$ is the feature representation in the embedding space.

4.3 Pre-trained Encoder

We learn an encoder by training an autoencoder network [Baldi, 2012]. An Encoder compresses the input at each time point into a latent-space representation, while Decoder reconstructs the input from the latent representation. After the encoder, $\mathcal{F}$, is trained, it can serve as a pre-trained representation learning model to project the original data into a latent embedding in Eq. (6). We provide more details on the model architecture in the arXiv version[2].

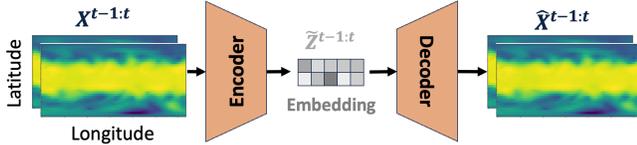$$\tilde{Z}^{t-1:t} = \mathcal{F}(X^{t-1}, X^t). \qquad (6)$$

Figure 3: Autoencoder structure.

4.4 Attention-based Denoiser Network

Our denoiser network consists of two blocks: cross-attention and U-net (as shown in Figure 4). Cross-attention mechanism [Hertz *et al.*, 2022] is employed to capture how past observations can contribute to the generation of future states. The embedding of past observations, $\tilde{Z}^{t-1:t}$, and the noise data X_n^{t+1} at diffusion step n, are projected to the same hidden dimension d with the following transformation:

$$Q = W_q \cdot X_n^{t+1}, K = W_k \cdot \tilde{Z}^{t-1:t}, V = W_v \cdot \tilde{Z}^{t-1:t}, \quad (7)$$

where $X_n^{t+1} \in \mathbb{R}^{(H \times W) \times C}$ and $\tilde{Z}^{t-1:t} \in \mathbb{R}^{(H \times W) \times d_z}$. $W_q \in \mathbb{R}^{d \times C}, W_k \in \mathbb{R}^{d \times d_z}, W_v \in \mathbb{R}^{d \times d_z}$ are

---

[2]Preprint version: https://arxiv.org/abs/2409.05975

learnable projection matrices. Then we implement the cross-attention mechanism by `Attention(Q, K, V) = softmax(`$\frac{QK^T}{\sqrt{d}}$`)V`. A visual depiction of the cross-attention mechanism is in the arXiv version.
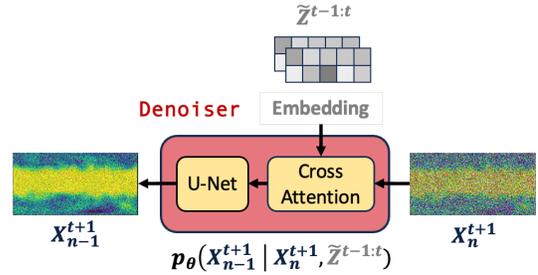

Figure 4: Attention-based denoiser structure.

U-Net [Ronneberger *et al.*, 2015] is utilized to recover the data by removing the noise added at each diffusion step. The *skip connection* technique in U-Net concatenates feature maps from the encoder to the corresponding decoder layers, allowing the network to retain fine-grained information that might be lost during downsampling. The detailed U-Net architecture is presented in the arXiv version.

4.5 Training Process

The training procedure is shown in the arXiv version.Firstly, we pre-train an encoder to learn the condition embedding of the past observations. Subsequently, we inject it into our conditional diffusion model and train CoDiCast with the devised loss function:

$$\mathcal{L}_{cond}(\theta) = \mathbb{E}_{X_0, \epsilon, n} \left\| \epsilon - \epsilon_\theta \left(X_n^{t+1}, n, \texttt{cond} \right) \right\|^2, \quad (8)$$

where $X_n^{t+1} = \sqrt{\bar{\alpha}_n} X_0^{t+1} + \sqrt{1 - \bar{\alpha}_n} \epsilon$, $\texttt{cond} = \mathcal{F}(X^{t-1:t})$, and ϵ_θ is the denoiser in Figure 4.

4.6 Inference Process

We first extract the conditional embedding representations, $\tilde{Z}^{t-1:t}$, by the pre-trained encoder, and then randomly generate a noise vector $X_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The sampled noise vector, X_N, is autoregressively denoised along the reversed chain to predict the target until n equals 1, we obtain the weather prediction $\hat{X}_0$ at the time $t+1$. Later, multi-step prediction can be implemented autoregressively - the output from the previous time step is the input while predicting the next step, as shown in Eq. (5). We provide the pseudocode in the arXiv version.

4.7 Ensemble Forecast

To enhance the reliability of weather forecasts, *ensemble forecast* strategy is often employed to capture the variability among forecasts by separately running multiple deterministic models, e.g., ensemble forecast suite (ENS) [Buizza, 2008]. In our approach, since `CoDiCast` is a probabilistic model that can generate a distribution of future weather scenarios rather than a single prediction, following [Price *et al.*, May 2024], we run the trained `CoDiCast` multiple times to get the ensemble instead. More specifically, by integrating both initial conditions and noise sampled from a Gaussian distribution, `CoDiCast` implements the ensemble forecast through multiple stochastic samplings during inference, capturing a range of possible forecasts for the uncertainty quantification.

5 Experiments

5.1 Dataset and Baselines

Dataset. ERA5 [Hersbach *et al.*, 2020] is a publicly available atmospheric reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Following the existing work [Verma *et al.*, 2024], we use the preprocessed $5.625°$ resolution (32×64) and 6-hour increment ERA5 dataset from WeatherBench [Rasp *et al.*, 2020]. We downloaded 5 variables for the globe: geopotential at 500 hPa pressure level (`Z500`), atmospheric temperature at 850 hPa pressure level (`T850`), ground temperature (`T2m`), 10 meter U wind component (`U10`) and 10 meter V wind component (`V10`). More details are in the arXiv version.

Baselines. We comprise the following baselines: ClimODE [Verma *et al.*, 2024]: a spatiotemporal continuous-time model that incorporates the physical knowledge of atmospheric *advection* over time; ClimaX [Nguyen *et al.*, 2023]: a state-of-the-art vision Transformer-based method trained on the same dataset (without pre-training that is used in the original paper); FourCastNet [Pathak *et al.*, 2022]: a global data-driven weather model using adaptive Fourier neural operators; Neural ODE [Chen *et al.*, 2018]: an ODE network that learns the time derivatives as neural networks by solving an ordinary differential equation; Integrated Forecasting System IFS [Rasp *et al.*, 2020]: a global numerical weather prediction (NWP) system, integrating multiple advanced physics-based models to deal with more meteorological variables across multiple altitudes. Our study focuses solely on a subset of these variables due to the limited computational resources, with IFS serving as the gold standard. For a fair comparison, all ML models use the same data set described in Section 5.1.

5.2 Experiments Design

We use data between 2006 and 2015 as the training set, data in 2016 as the validation set, and data between 2017 and 2018 as the testing set. We assess the global weather forecasting capabilities of our method `CoDiCast` by predicting the weather at a future time $t + \Delta t$ ($\Delta t = 6$ to 144 hours) based on the past two time units. To quantify the uncertainty in weather prediction, we generate an "ensemble" forecast by running `CoDiCast` five times during the inference phase.

Training. We first pretrain an `encoder` model with the `Autoencoder` architecture. For the diffusion model, we used U-Net as the denoiser network with 1000 diffusion/denoising steps. The architecture is similar to that of DDPM [Ho *et al.*, 2020] work. We employ four U-Net units for both the downsampling and upsampling processes. Before training, we apply Max-Min normalization [Ali *et al.*, 2014] to scale the input data within the range $[0, 1]$, mitigating potential biases stemming from varying scales [Shi *et al.*, 2023]. `Adam` was used as the optimizer, where the learning rate $= 2e^{-4}$, decay steps $= 10000$, decay rate $= 0.95$. The batch size and number of epochs were set to 64 and 800, respectively. More training details and model configurations are in the arXiv version. We conduct all experiments on one NVIDIA A100 GPU with 80GB memory.

Evaluation Metrics. Following [Verma *et al.*, 2024], we use latitude-weighted Root Mean Square Error (RMSE) and Anomaly Correlation Coefficient (ACC) as deterministic metrics. RMSE measures the average difference between values predicted by a model and the actual values. ACC is the correlation between prediction anomalies relative to climatology and ground truth anomalies relative to climatology. It is a critical metric in climate science to evaluate the model's performance in capturing unusual weather or climate events. The formulas of these metrics are in the arXiv version.

5.3 Quantitative Evaluation

Accuracy. We compare different models in forecasting five primary meteorological variables as described in Section 5.1. Table 1 shows that `CoDiCast` presents superior performance over other MLWP baselines, demonstrating diffusion models can capture the weather dynamics and make predictions accurately. However, there is still room for `CoDiCast` to be improved compared with the gold-standard `IFS` model, which integrates more meteorological variables.

Uncertainty. The gray-shaded error range in Table 1 represents model uncertainty, with fluctuations remaining within 10% of the ground truth scale. This indicates the robustness of ML models. However, `ForeCastNet` predictions fluctuate relatively larger due to the sensitive selection of the dropout rate. We provide a case study of `CoDiCast` forecast for 72 hours with uncertainty quantification in Figure 5. It shows the mean prediction tracks the general trend of the ground truth and the uncertainty grows as the lead time increases. Besides, we can tell that most actual values fall within the 1 or 2 standard deviations (σ) ranges, indicating these predictions are reasonably accurate.

Variable	Lead time (Hours)	RMSE (↓)						ACC (↑)					
		NODE	ClimaX	ForeCastNet	ClimODE	CoDiCast	IFS	NODE	ClimaX	ForeCastNet	ClimODE	CoDiCast	IFS
Z500	6	300.6	247.5	222.7±18.1	102.9±9.3	**73.1**±6.7	26.9	0.96	0.97	0.97	0.99	**0.99**	1.00
	12	460.2	265.3	310.9±22.7	134.8±12.3	**114.2**±8.9	33.8	0.88	0.96	0.95	0.99	**0.99**	0.99
	24	877.8	364.9	402.6±27.3	193.4±16.3	**186.5**±11.8	51.0	0.70	0.93	0.92	0.98	**0.98**	0.99
	72	N/A	687.0	755.3±45.8	478.7±48.5	**451.6**±39.5	123.2	N/A	0.73	0.75	0.88	**0.92**	0.98
	144	N/A	801.9	956.1±59.1	783.6±37.3	**757.5**±42.8	398.7	N/A	0.58	0.64	0.61	**0.78**	0.86
T850	6	1.82	1.64	1.75±0.16	1.16±0.06	**1.02**±0.05	0.69	0.94	0.94	0.94	0.97	**0.99**	0.99
	12	2.32	1.77	2.15±0.20	1.32±0.13	**1.26**±0.10	0.75	0.85	0.93	0.92	0.96	**0.99**	0.99
	24	3.35	2.17	2.51±0.27	1.55±0.18	**1.52**±0.16	0.87	0.72	0.90	0.89	0.95	**0.97**	0.99
	72	N/A	3.17	3.69±0.34	2.58±0.16	**2.54**±0.14	1.15	N/A	0.76	0.77	0.85	**0.93**	0.96
	144	N/A	3.97	4.29±0.42	3.62±0.21	**3.61**±0.19	2.23	N/A	0.69	0.71	0.77	**0.85**	0.81
T2m	6	2.72	2.02	2.05±0.18	1.21±0.09	**0.95**±0.07	0.69	0.82	0.92	0.94	0.97	**0.99**	0.99
	12	3.16	2.26	2.49±0.21	1.45±0.10	**1.21**±0.07	0.77	0.68	0.90	0.92	0.96	**0.99**	0.99
	24	3.86	2.37	2.78±0.26	1.40±0.09	1.45±0.07	1.02	0.79	0.89	0.91	0.96	**0.99**	0.99
	72	N/A	2.87	3.77±0.32	2.75±0.49	**2.39**±0.37	1.26	N/A	0.83	0.85	0.85	**0.96**	0.96
	144	N/A	3.38	4.39±0.41	3.30±0.23	3.45±0.22	1.78	N/A	0.83	0.81	0.79	**0.91**	0.82
U10	6	2.30	1.58	1.98±0.17	1.41±0.07	**1.24**±0.06	0.61	0.85	0.92	0.87	0.91	**0.95**	0.98
	12	3.13	1.96	2.58±0.21	1.81±0.09	**1.50**±0.08	0.76	0.70	0.88	0.78	0.89	**0.93**	0.98
	24	4.10	2.49	3.02±0.27	2.01±0.10	**1.87**±0.09	1.11	0.50	0.80	0.71	0.87	**0.89**	0.97
	72	N/A	3.70	4.17±0.36	3.19±0.18	**3.15**±0.19	1.57	N/A	0.45	0.41	0.66	**0.71**	0.94
	144	N/A	4.24	4.63±0.45	4.02±0.12	4.25±0.15	3.04	N/A	0.30	0.28	0.35	**0.42**	0.72
V10	6	2.58	1.60	2.16±0.19	1.53±0.08	**1.30**±0.06	0.61	0.81	0.92	0.86	0.92	**0.95**	1.00
	12	3.19	1.97	2.73±0.23	1.81±0.12	**1.56**±0.09	0.79	0.61	0.88	0.76	0.89	**0.93**	0.99
	24	4.07	2.48	3.15±0.28	2.04±0.10	**1.94**±0.14	1.33	0.35	0.80	0.68	0.86	**0.89**	1.00
	72	N/A	3.80	4.26±0.34	3.30±0.22	**3.18**±0.19	1.67	N/A	0.39	0.38	0.63	**0.68**	0.93
	144	N/A	4.42	4.64±0.45	4.24±0.10	**4.21**±0.18	3.26	N/A	0.25	0.27	0.32	**0.37**	0.71

Table 1: Latitude-weighted RMSE (↓) and ACC (↑) comparison on global weather forecasting. The results of NODE, ClimaX, and ClimODE models are from the ClimODE paper. N/A represents the values that are not available from their paper. ForeCastNet was re-trained with their code. We employ Monte Carlo Dropout during the inference to compute the uncertainty. We mark the scores in **bold** if our model `CoDiCast` performs the best among MLWP methods.

Inference efficiency. Generally, numerical weather prediction models (e.g., `IFS`) require around 50 minutes for the medium-range global forecast, while deterministic ML weather prediction models take less than 1 minute [Rasp *et al.*, 2020] but cannot model the weather uncertainty. `CoDiCast` needs about 3.6 minutes (see the last row in Table 2) for the global weather forecast, potentially balancing the efficiency and accuracy with essential uncertainty quantification. The efficiency also depends on the model complexity.

5.4 Qualitative Evaluation

In Figure 6, we qualitatively evaluate the performance of `CoDiCast` on global forecasting tasks for all target variables, `Z500`, `T850`, `T2m`, `U10` and `V10` at the lead time of 6 hours. The first row is the ground truth of the target variable, the second row is the prediction and the last row is the difference between the model prediction and the ground truth. From the scale of color bars, we can tell that the error percentage is less than 3% for variables `Z500`, `T850`, and `T2m`. However, error percentages over 50% exist for `U10` and `V10` even though only a few of them exist. We also observe that most higher errors appear in the high-latitude ocean areas, probably due to the sparse data nearby. We provide visualizations for longer lead times in the arXiv version.

5.5 Ablation Study

`CoDiCast` includes two important components: *pre-trained encoder* and *cross attention*. To study their effectiveness, we conduct an ablation study: (a) **No-encoder** directly considers past observations as conditions to the diffusion model; (b) **No-cross-attention** simply concatenates the embedding

and the noisy sample at each denoising step; (c) **No-encoder-cross-attention** concatenates the past observations and the noisy sample at each denoising step. Figure 7 shows that the full version of `CoDiCast` consistently outperforms all other variants, verifying positive contributions of both components.
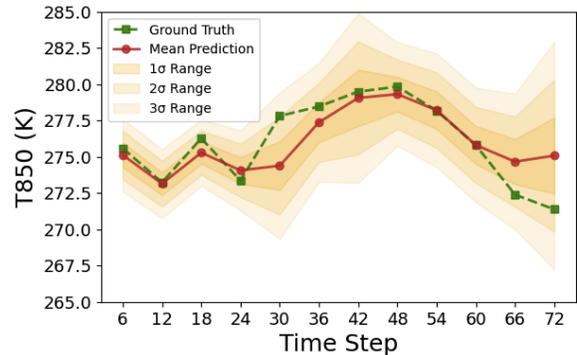

Figure 5: Model Forecasts with confidence intervals.

5.6 Parameter Study

Table 2 shows that the accuracy improves as the number of diffusion steps, N, increases when $N < 1000$, indicating that more intermediate steps are more effective in learning the imperceptible attributes during the denoising process. However, when $1000 < N < 2000$, the accuracy remains approximately flat but the inference time keeps increasing linearly. Considering the trade-off between accuracy and efficiency, we finally set $N = 1000$. Additionally, we study the effect of "linear" and "quadratic" variance scheduling β, where $\beta \in [0.0001, 0.02]$. Figure 8 shows that the "linear" variance scheduling performs better than the "quadratic" one.
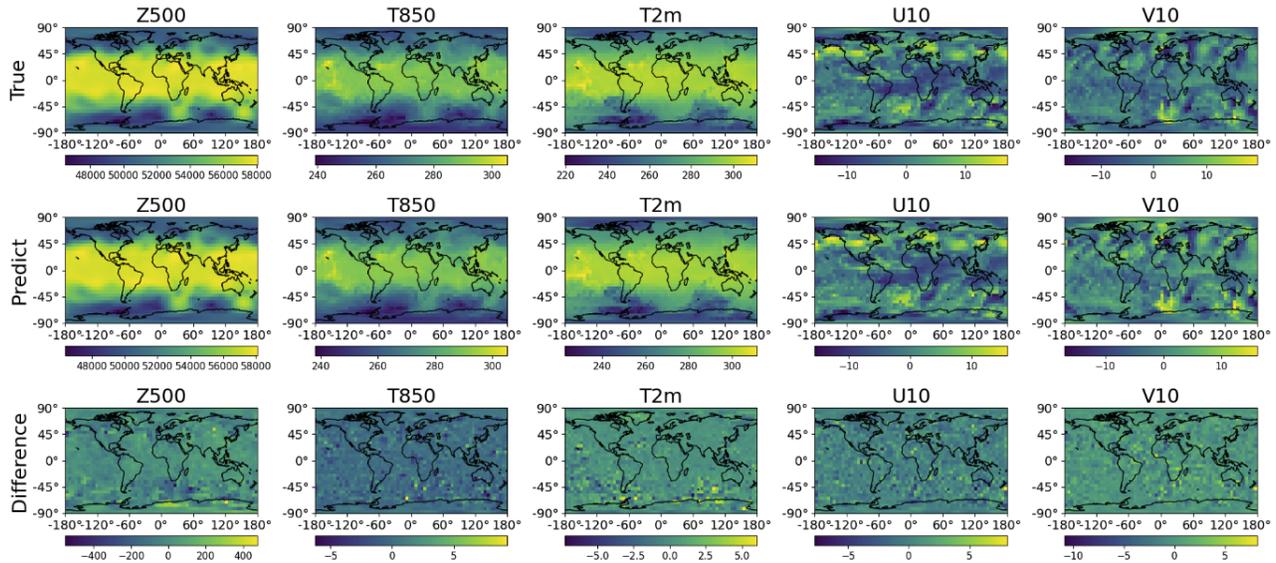
Figure 6: Visualization of true and predicted values across five meteorological variables at 6 hours lead time.
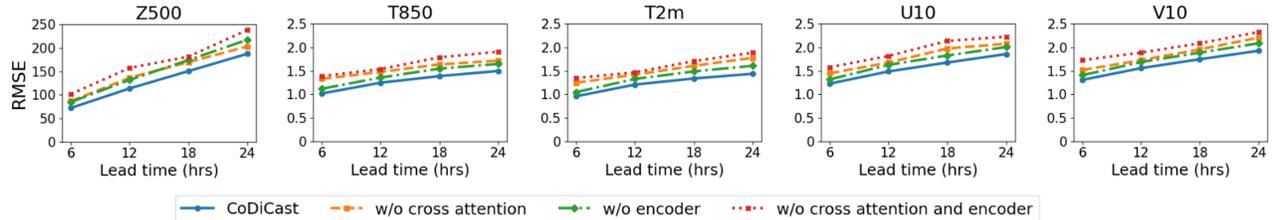

Figure 7: Ablation study to study the effect of pre-trained `encoder` and `cross-attention` components.
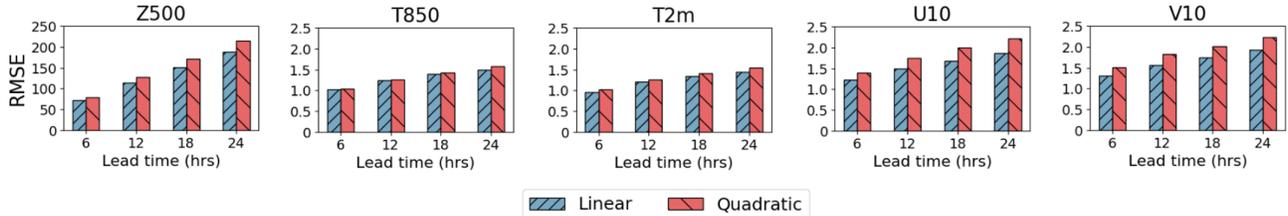

Figure 8: Effect of linear and quadratic variance scheduling methods.

Variable	Lead Time (Hours)	Diffusion Step					
		250	500	750	1000	1500	2000
Z500	24	696.1	324.8	190.6	**186.5**	193.5	191.9
T850	24	3.88	2.38	1.53	**1.52**	1.56	1.58
T2m	24	5.26	2.79	1.63	**1.44**	1.50	1.53
U10	24	2.74	2.05	**1.81**	1.87	1.99	2.01
V10	24	2.43	2.11	**1.89**	1.94	2.04	2.06
Inference time (min)		~ 1.1	~ 1.9	~ 2.8	~ 3.6	~ 6.5	~ 8.3

Table 2: RMSE with various diffusion steps at 24 hours lead time.

6 Conclusions

In this work, we start with analyzing the limitations of current deterministic numerical weather prediction (NWP) and machine-learning weather prediction (MLWP) approaches. They either require substantial computational costs or lack uncertainty quantification in their forecasts. We propose a conditional diffusion model, CoDiCast, to address these

limitations. Experimental results demonstrate it can **simultaneously** complete more accurate predictions than existing MLWP-based models and a faster inference than NWP-based models while being capable of providing uncertainty quantification compared to deterministic methods. In conclusion, our model offers three key characteristics at the same time.

7 Collaborations and Broader Impacts

Machine learning (ML) and atmospheric researchers were involved in the multidisciplinary collaborations. The domain experts provide ML researchers with data sources and contextual experiment results while obtaining insights into the potential applications of diffusion models in atmospheric sciences. *Impact:* Our work facilitates weather warnings such as extreme heat and cold, and storm warnings, which could help the government, local agencies, and residents to take necessary preventive and protective measures to minimize losses.

Acknowledgments

References

[Ali *et al.*, 2014] Peshawa Jamal Muhammad Ali, Rezhna Hassan Faraj, Erbil Koya, Peshawa J Muhammad Ali, and Rezhna H Faraj. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1):1–6, 2014.

[Asperti *et al.*, 2023] Andrea Asperti, Fabio Merizzi, Alberto Paparella, Giorgio Pedrazzi, Matteo Angelinelli, and Stefano Colamonaco. Precipitation nowcasting with generative diffusion models. *arXiv preprint arXiv:2308.06733*, 2023.

[Balaji *et al.*, 2022] V Balaji, Fleur Couvreux, Julie Deshayes, Jacques Gautrais, Frédéric Hourdin, and Catherine Rio. Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, 119(47):e2202075119, 2022.

[Baldi, 2012] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49, Edinburgh, Scotland, 2012. JMLR Workshop and Conference Proceedings.

[Bauer *et al.*, 2015] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[Ben Bouallègue *et al.*, 2024] Zied Ben Bouallègue, Mariana CA Clare, Linus Magnusson, Estibaliz Gascon, Michael Maier-Gerber, Martin Janoušek, Mark Rodwell, Florian Pinault, Jesper S Dramsch, Simon TK Lang, et al. The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024.

[Bi *et al.*, 2023] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[Brempong *et al.*, 2022] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022.

[Buizza, 2008] Roberto Buizza. Comparison of a 51-member low-resolution (t l 399l162) ensemble with a 6-member high-resolution (t l 799l91) lagged-forecast ensemble. *Monthly weather review*, 136(9):3343–3362, 2008.

[Bülte *et al.*, 2024] Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *arXiv preprint arXiv:2403.13458*, 2024.

[Chen *et al.*, 2018] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[Chen *et al.*, 2023a] Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

[Chen *et al.*, 2023b] Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

[ECMWF, 2023] ECMWF. Medium-range forecasts, 2023.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[Gao *et al.*, 2024] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hersbach *et al.*, 2020] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[Hewage *et al.*, 2021] Pradeep Hewage, Marcello Trovati, Ella Pereira, and Ardhendu Behera. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, 24(1):343–366, 2021.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jaseena and Kovoor, 2022] KU Jaseena and Binsu C Kovoor. Deterministic weather forecasting models based on intelligent predictors: A survey. *Journal of king saud university-computer and information sciences*, 34(6):3393–3412, 2022.

[Keisler, 2022] Ryan Keisler. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575*, 2022.

[Kochkov *et al.*, 2024] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, pages 1–7, 2024.

[Lam *et al.*, 2023] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[Leinonen *et al.*, 2023] Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

[Li *et al.*, 2022] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[Luo and Hu, 2021] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2837–2845, 2021.

[Morley *et al.*, 2018] SK Morley, DT Welling, and JR Woodroffe. Perturbed input ensemble modeling with the space weather modeling framework. *Space Weather*, 16(9):1330–1347, 2018.

[Nguyen *et al.*, 2023] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[Nichol *et al.*, 2021] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[Palmer, 2012] TN Palmer. Towards the probabilistic earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 138(665):841–861, 2012.

[Palmer, 2019] Tim Palmer. The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.

[Pathak *et al.*, 2022] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[Price *et al.*, May 2024] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Timo Ewalds, Andrew El-Kadi, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, May 2024.

[Rasp *et al.*, 2020] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[Rodwell and Palmer, 2007] MJ Rodwell and TN Palmer. Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 133(622):129–146, 2007.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, October 5-9, 2015, proceedings, part III 18*, pages 234–241, 2015.

[Shi *et al.*, 2023] Jimeng Shi, Rukmangadh Myana, Vitalii Stebliankin, Azam Shirali, and Giri Narasimhan. Explainable parallel rcnn with novel feature representation for time series forecasting. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 56–75, Torino, Italy, 2023. Springer.

[Shi *et al.*, 2025] Jimeng Shi, Zeda Yin, Arturo Leon, Jayantha Obeysekera, and Giri Narasimhan. Fidlar: Forecast-informed deep learning architecture for flood mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28377–28385, 2025.

[Siddique *et al.*, 2022] Talha Siddique, Md Shaad Mahmud, Amy M Keesee, Chigomezyo M Ngwira, and Hyunju Connor. A survey of uncertainty quantification in machine learning for space weather prediction. *Geosciences*, 12(1):27, 2022.

[Slingo and Palmer, 2011] Julia Slingo and Tim Palmer. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4751–4767, 2011.

[Verma *et al.*, 2024] Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. *arXiv preprint arXiv:2404.10024*, 2024.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.