Deep Learning and Foundation Models for Weather Prediction: A Survey

Jimeng Shi[1], **Azam Shirali**[2], **Bowen Jin**[1], **Sizhe Zhou**[1], **Wei Hu**[1], **Rahuul Rangaraj**[2]
Zhaonan Wang[3], **Yanzhao Wu**[2], **Leonardo Bobadilla**[2], **Upmanu Lall**[4,5],
Shaowen Wang[1], **Jiawei Han**[1], **Giri Narasimhan**[2]
[1]University of Illinois Urbana-Champaign,
[2]Florida International University
[3]New York University Shanghai,
[4]Arizona State University,
[5]Columbia University
{jimeng8, shaowen}@illinois.edu, giri@fiu.edu

Abstract

Numerical weather prediction (NWP) models remain the cornerstone of atmospheric sciences. Yet, deep learning (DL) is challenging this paradigm by its ability to capture intricate spatio-temporal patterns and deliver ultra-fast predictions. Analogous to the foundation models (e.g., ChatGPT) in natural language processing, foundation models in the weather/climate domain have also been developed. This paper reviews DL and foundation models for weather prediction by highlighting their strengths and limitations. In particular, we carefully examine them from the perspective of their training paradigms: deterministic predictive learning, probabilistic generative learning, and pre-training & fine-tuning. For each paradigm, we summarize the underlying model architectures, training methods, and respective features. To facilitate further study, we provide a curated repository featuring categorized papers, open-source code, and benchmark datasets. Finally, we discuss and suggest potential research directions across new tasks and models in weather data storage and management, and operational deployment, further inspiring innovations in this rapidly evolving field. GitHub: https://github.com/JimengShi/DL-Foundation-Models-Weather.

1 Introduction

Accurate and timely weather prediction is critical to mitigate the impacts of extreme weather events and support decision-making in sectors such as agriculture, transportation, and disaster management [Abbass *et al.*, 2022]. Physics-based numerical weather prediction (NWP) models have remained the cornerstone of the field for a long time. These models predict weather by numerically solving differential equations that govern complex atmospheric, terrestrial, and oceanic interactions [Nguyen *et al.*, 2023]. Despite significant advances, they face notable challenges: computationally slow, simplified assumptions about atmospheric dynamics, uncertainty quantification [Abbass *et al.*, 2022].

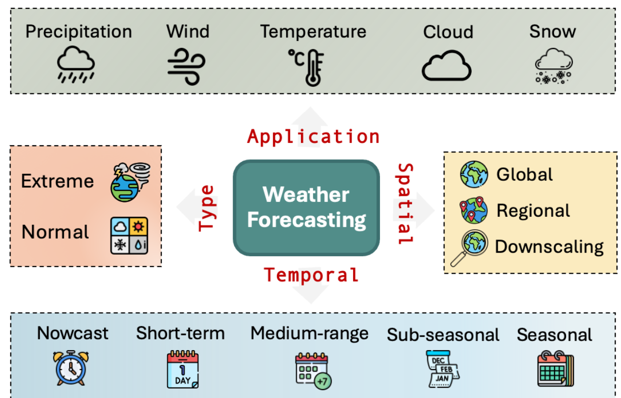

Figure 1: Perspectives of weather prediction.

Data-driven models have been increasingly applied to weather and climate modeling, demonstrating remarkable advances in precision, computational efficiency, and uncertainty quantification. They can capture complex atmospheric dynamics in an end-to-end fashion, eliminating the reliance on explicit prior knowledge of physical relationships. Deterministic models such as `Pangu` [Bi *et al.*, 2023] and `GraphCast` [Lam *et al.*, 2022] have achieved state-of-the-art performance in medium-range (10-day) global weather prediction, matching or even surpassing traditional NWP methods in terms of accuracy on the ERA5 benchmark dataset while dramatically reducing computational costs (up to three orders of magnitude). However, their predictions are often blurry since they are trained by minimizing point-wise loss functions. To overcome this limitation, probabilistic models have been studied for weather prediction while achieving uncertainty quantification for model predictions. They consider the prediction process as probabilistic sampling (i.e., generation) conditioning on necessary constraints. `CasCast` [Gong *et al.*, 2024] and `Gencast` [Price *et al.*, 2025] leverage diffusion models for probabilistic precipitation nowcasting and weather prediction. More recently, foundation models have gained traction in climate and weather modeling as an emerging paradigm [Bodnar *et al.*, 2025].

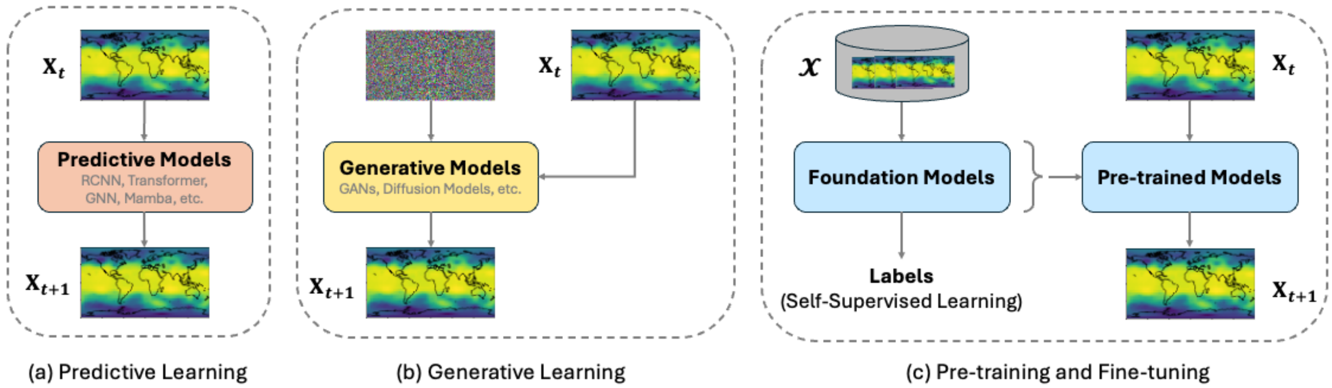

Figure 2: Various training paradigms for weather prediction. For clarity, this visualization focuses on single-step predictions for a single variable.

These models are pre-trained on massive historical weather datasets to learn generalizable and comprehensive knowledge, which can then be fine-tuned for diverse downstream tasks (e.g., weather prediction and climate downscaling).

With the rapid development of data-driven (e.g., deep learning) models in weather and climate science, a systematic and up-to-date survey is essential for consolidating knowledge and guiding future research. In this work, we propose a systematic taxonomy of existing DL models for weather prediction based on their training paradigms: *deterministic predictive learning*, *probabilistic generative learning*, and *pre-training and fine-tuning*, as presented in Fig. 2. Building on this framework, we provide a comprehensive survey of state-of-the-art models, critically analyzing their strengths, limitations, and applicability to various prediction tasks. Our contributions are summarized as follows:

- **Novel Taxonomy.** We introduce a systematic categorization of existing DL models for weather prediction based on their training paradigms: deterministic predictive learning, probabilistic generative learning, and pre-training and fine-tuning.
- **Comprehensive Overview.** We present a detailed and up-to-date survey of the state-of-the-art models, analyzing their strengths, limitations, and applications in weather prediction.
- **Extensive Resources.** We compile an extensive repository of resources, including benchmark datasets, open-source codes, and real-world applications to support future research.
- **Future Directions.** We also outline a forward-looking roadmap, highlighting six critical research directions advancing AI applications for weather science.

2 Background

Fig. 1 presents weather prediction tasks in four aspects. (1) *Temporal*: prediction for future time point(s) of interest, $t + \Delta t$, where the lead time of prediction $\Delta t \approx \{$minutes, hours, days, weeks, months, years$\}$. Based on the prediction range, it covers nowcast (minutes to hours), medium-range prediction (days to two weeks), sub-seasonal, and seasonal prediction (over months to years). (2) *Spatial*: prediction for global and regional areas. (3) *Applications*: focus on predicting weather variables of interest. (4) *Event Type*: predictions may be for extreme events, such as heatwaves, snowstorms, hurricanes, tropical cyclones, and heavy rainfall.

3 Overview and Taxonomy

This survey is structured across three dimensions: modeling paradigm, model architecture (backbone), and spatial scope. First, we distinguish various methods based on their modeling/learning paradigms: deterministic predictive learning, probabilistic generative learning, and pre-training and fine-tuning (see Fig. 2). Second, a taxonomy is provided in Fig. 3, unfolded with model backbones, such as RNNs, Transformers, GNNs, Mamba, GANs, and Diffusion Models. Third, spatial coverage is discussed with global vs. domain scope.

3.1 Predictive Learning-based Models

Predictive learning methods are usually *deterministic*, where models aim to predict future states of weather variables (e.g., temperature, humidity, wind speed, and precipitation) based on past and current observations. These models are typically built to recognize weather patterns or dependencies in historical data by minimizing a point-wised loss function (e.g., mean absolute errors). We systematically categorize these predictive models into global general-purpose models and domain-specific models. Each categorization is further discussed with various model architectures.

Global General-Purpose Models

They are often designed for global multi-variable prediction and typically operate at coarse spatial resolutions ($0.25^\circ \sim 5.625^\circ$) and temporal intervals (6 or 12 hours).

Transformer-based models. Transformer [Vaswani *et al.*, 2017] is widely used for sequential modeling with the powerful *attention* mechanism. Vision Transformer (ViT) is a variant of Transformer used for image-style data. In the weather field, it treats a global weather map as an image and processes it by dividing the input into a sequence of small, non-overlapping patches. This patch-based approach enables the model's self-attention mechanism to learn complex,

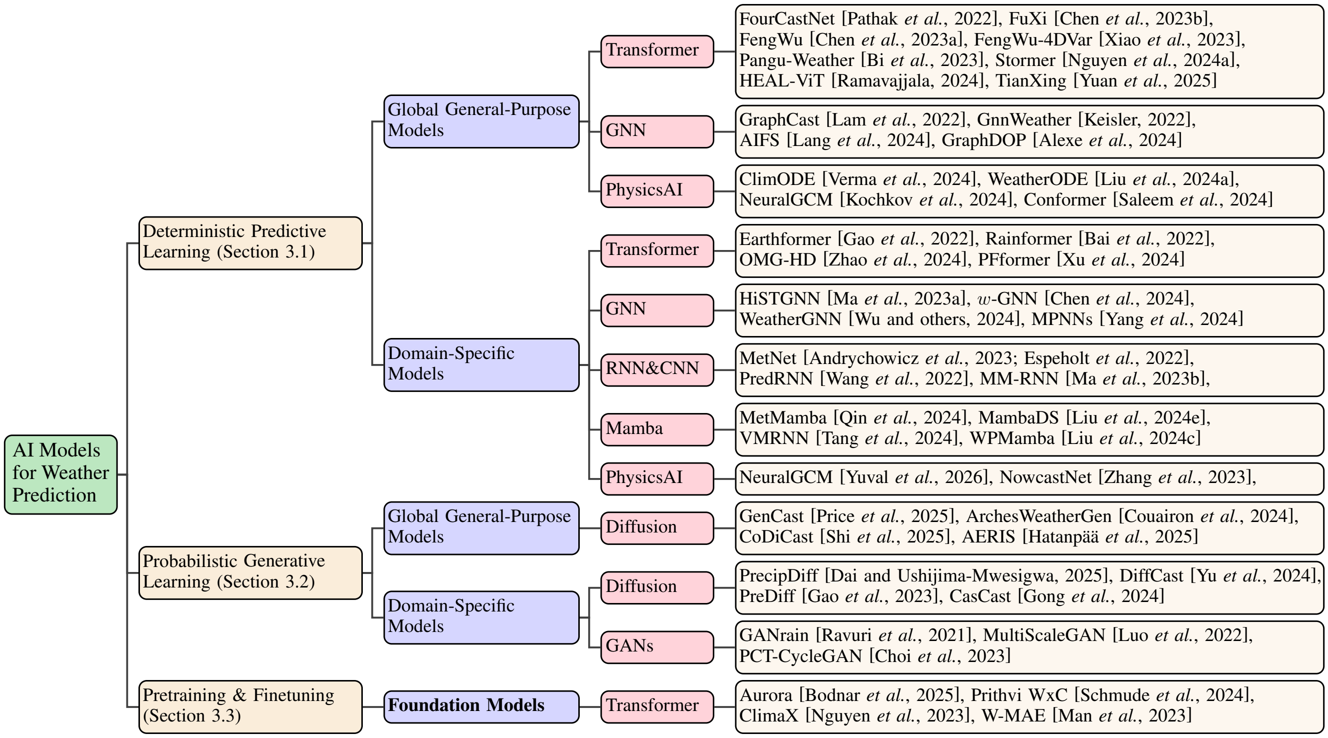

Figure 3: A comprehensive taxonomy of deep learning and foundation models for weather prediction from the perspectives of training paradigms (dark yellow), model scopes (purple), and model architectures (pink). **Note:** Due to space constraints, this review focuses on a representative selection of recent models. We acknowledge that a broader body of related work exists beyond the scope of this review.

long-range dependencies between weather patterns globally. Built on the ViT architecture, `FourCastNet` [Pathak *et al.*, 2022] and `HEAL-ViT` [Ramavajjala, 2024] are developed for global weather prediction and achieve comparable performance with NWP models. Specifically, `FourCastNet` also employs a Fourier transform-based token-mixing scheme [Guibas *et al.*, 2021], while `HEAL-ViT` uses ViT on a spherical mesh, benefiting from both the spatial homogeneity enjoyed by graph-based models and efficient attention-based mechanisms exploited by transformers. Moreover, `TianXing` [Yuan *et al.*, 2025] proposes a variant attention mechanism with *linear* complexity for global weather prediction, significantly diminishing GPU resource demands, with only a marginal compromise in accuracy.

However, the above methods do not directly learn the interdependencies among atmospheric variables. `FengWu` [Chen *et al.*, 2023a] introduces a multi-encoder design that processes each meteorological variable independently, followed by a transformer-based fusion network that captures inter-variable dependencies. This design aims to preserve variable-specific dynamics while learning complex inter-variable relationships. `FengWu-4DVar` [Xiao *et al.*, 2023] extends `FengWu` with the Four-Dimensional Variational (4DVar) assimilation algorithm, accomplishing both weather prediction and data assimilation.

Despite impressive performance, these models generate multi-step forecasts in an autoregressive manner, being prone to accumulating errors as the length of the prediction window increases. To mitigate this phenomenon, `Pangu-Weather` [Bi *et al.*, 2023] and `FuXi` models [Chen *et al.*, 2023b] offer a solution that uses a hierarchical temporal aggregation algorithm at the inference phase. Given a forecast goal with a certain lead time, the greedy algorithm is used to guarantee the minimal number of iterations of the trained models for that forecast window. For example, for a 7-day forecast, `Pangu` executes the 24-hour forecast 7 times, while for a 23-hour forecast, `Pangu` executes the 6-hour forecast 3 times, followed by the 3-hour forecast 1 time and the 1-hour forecast 2 times.

Beyond the direct and iterative prediction, the `Stormer` model [Nguyen *et al.*, 2024a] needs the explicit time point, $t + \Delta t$ to guide the models for predictions.

GNN-based models. Graph Neural Networks (GNNs) have been applied for global weather prediction by representing the Earth's coordinates (i.e., latitude, longitude, altitude) as a graph [Keisler, 2022]. By modeling the Earth as a graph with nodes representing spatial locations and edges encoding their relationships, such models capture spatial dependencies in weather patterns. GNN-based methods effectively integrate local and global weather dynamics by modeling spatial locations and their relationships with nodes and edges in a graph. Building on this, `GraphCast` [Lam *et al.*, 2022] expanded these capabilities to forecast hundreds of weather variables at a high 0.25° resolution with a 10-day lead time. Trained on ERA5 reanalysis data [Rasp *et al.*, 2023], `GraphCast` demonstrates superior performance in predicting severe weather events compared to the ECMWF's High-RESolution forecast (HRES). More recently, the ECMWF introduced its own GNN-based architectures: `AIFS` [Lang *et al.*, 2024] and `GraphDOP` [Alexe *et al.*, 2024]. Notably, `GraphDOP` operates entirely within observation space, bypassing the need for gridded climatology or NWP reanalysis feedback.

GNNs are particularly well-suited for atmospheric modeling due to their capacity to represent irregular spatial structures, capture non-local dependencies, and generalize across heterogeneous domains. However, they face significant challenges. First, defining graph topologies that faithfully represent fluid dynamics is nontrivial; static graphs may fail to capture evolving multi-scale spatiotemporal interactions. Second, GNNs often struggle with long-range temporal dependencies unless augmented with recurrent modules or attention mechanisms. Finally, scaling message-passing operations to high-resolution global datasets remains computationally demanding, particularly as the graph size increases.

Physics-AI-based models. Data-driven methods often operate as black-box models, complicating verification and operational deployment. Therefore, hybrid frameworks that integrate machine learning with physical principles have been explored to enhance interpretability and reliability. `ClimODE` [Verma *et al.*, 2024] integrates the principle of *advection* to model weather as a spatiotemporal continuous-time process. By utilizing continuity Ordinary Differential Equations (ODEs), it captures value-conserving dynamics and learns global weather transport as a neural flow. Building on this, `WeatherODE` [Liu *et al.*, 2024a] incorporates wave equation theory and a time-dependent source model to achieve higher accuracy in solving advection equations. Moving toward more comprehensive physics integration, `NeuralGCM` [Kochkov *et al.*, 2024] employs a differentiable dynamical core to solve a broader suite of primitive equations, including momentum, thermodynamics, and hydrostatic approximations. This model couples these equations with a learned physics module that parameterizes subgrid-scale processes, such as cloud formation and radiative transport, that are traditionally difficult to resolve.

By explicitly incorporating governing equations and physical constraints, these hybrid models significantly improve physical consistency and conservation law adherence. Such frameworks are particularly robust in data-sparse or noisy regimes, where physical knowledge acts as a critical regularizer for the learning process.

Domain-Specific Models

We present domain-specific predictive models for regional or single-variable weather predictions.

Transformer-based models. `Earthformer` [Gao *et al.*, 2022] proposes a generic, flexible, and efficient space-time attention block (Cuboid Attention) for regional precipitation prediction. `Rainformer` [Bai *et al.*, 2022] combines CNN and Swin Transformer for precipitation nowcasting. `PFformer` [Xu *et al.*, 2024] utilizes an inverted Transformer [Liu *et al.*, 2024d] to learn spatial dependencies among multiple observation stations for short-term precipitation prediction. for extreme precipitation nowcasting. The `OMG-HD` model [Zhao *et al.*, 2024] leverages the Swim Transformer for regional high-resolution weather prediction trained with multiple observational data, e.g., stations, radar, and satellite.

GNN-based models. `HiSTGNN` [Ma *et al.*, 2023a] incorporates an adaptive graph learning module comprising a global graph representing regions and a local graph capturing meteorological variables for each region. The w`-GNN` model [Chen *et al.*, 2024] leverages Graph Neural Networks coupled with physical factors for precipitation forecast in China. `WeatherGNN` [Wu and others, 2024] proposes a fast hierarchical Graph Neural Network (FHGNN) to extract the spatial dependencies. The `MPNN` model [Yang *et al.*, 2024] exploits heterogeneous GNNs for both station-observed and gridded weather data, where the node at the prediction location aggregates information from its heterogeneous neighboring nodes by message passing.

RNN- & CNN-based models. `MetNet` [Espeholt *et al.*, 2022] leverages CNNs and LSTMs for precipitation prediction for lead times of 8 and 12 hours. `MetNet-3` [Andrychowicz *et al.*, 2023] significantly extends both the lead times (up to 24 hours) and variables (precipitation, wind, temperature). `MM-RNN` [Ma *et al.*, 2023b] introduces knowledge of elements to guide precipitation prediction and learn the underlying atmospheric motion laws using RNNs. Earlier stage, a representative work includes but is not limited to `Predrnn++` [Wang *et al.*, 2018].

Mamba-based models. `MetMamba` [Qin *et al.*, 2024] exploits Mamba's selective scan to achieve token (spatial, temporal) mixing and channel mixing to capture more complex spatiotemporal dependencies in weather data. `MambaDS` [Liu *et al.*, 2024e] attempts to use the selective state space model (Mamba) for the meteorological field downscaling. `VMRNN` [Tang *et al.*, 2024] develops an innovative architecture tailored for spatiotemporal prediction by integrating Vision Mamba and LSTM, surpassing established vision models in both efficiency and accuracy.

We observed that `Mamba` [Gu and Dao, 2023] has thus far been applied primarily to domain-specific tasks. Given its high efficiency in modeling long-range dependencies with linear computational complexity, it is potentially promising for extension to global-scale weather prediction.

Physics-AI-based models. `NowcastNet` [Zhang *et al.*, 2023] is a nowcasting model for extreme precipitation that unifies both physical first principles and statistical-learning methods. `PhysicsAI` [Das *et al.*, 2024] evaluated `NowcastNet` model with a case study on the Tennessee Valley Authority (TVA) service area, and found superior performance than the High Resolution Rapid Refresh (HRRR) model. Recently, `NeuralGCM` [Yuval *et al.*, 2026] was directly trained on satellite-based precipitation observations, unlike previous attempts at hybrid models that relied on high-resolution simulations as training data. Surprisingly, it outperforms the existing global circulation models (GCMs), the ERA5 reanalysis, and a global cloud-resolving model in simulating precipitation.

3.2 Generative Learning-based Models

Generative models can be used for weather *prediction* by treating them as *generative* processes conditioned on observations from the past. More significantly, since these generative models are probabilistic, they are well-suited to generate ensemble forecasts that help quantify the uncertainty in predictions, facilitating informed decision-making.

General-Purpose Large Models

Diffusion-based models. Generative diffusion models have been developed and applied for probabilistic weather prediction. `GenCast` [Price *et al.*, 2025] employs diffusion models for probabilistic weather predictions conditioning on the past two observations, generating an ensemble of stochastic 15-day global forecasts, at 12-hour steps and $0.25°$ latitude-longitude resolution, for over 80 surface and atmospheric variables. `CoDiCast` [Shi *et al.*, 2025] leverages a *pre-trained* encoder to learn embeddings based on observations from the recent past and a *cross-attention* mechanism to guide the generation process to predict future weather states. During the inference phase, these three methods generate predictions at a single step and roll out for long-term output autoregressively. However, they accumulate errors for high temporal resolution due to the many rollout steps. Different from the aforementioned methods, `ArchesWeatherGen` [Couairon *et al.*, 2024] integrates deterministic and probabilistic models by first outputting deterministic forecasts, and then enhancing them with probabilistic modeling residuals (i.e., the differences between its predictions and ERA5 data) to generate ensemble forecasts.

The brightest point is that diffusion models can efficiently generate a diverse ensemble of potential future weather states, which is crucial for quantifying uncertainty. While earlier models were computationally intensive due to autoregressive generation, newer approaches are being developed to produce forecasts in a single step, improving efficiency and temporal consistency. `AERIS` [Hatanpää *et al.*, 2025] further advances the scaling efficiency at high resolutions using a window parallelism with sequence and pipeline parallelism to shard window-based transformers.

Domain-Specific Models

GAN-based models. `GANrain` [Ravuri *et al.*, 2021] employs a conditional generative adversarial network (GAN) for the precipitation prediction problem, where the generator generates future precipitation frames and the discriminator learns to distinguish whether a sample is coming from the original training data or was generated by the generator. `PCT-CycleGAN` [Choi *et al.*, 2023] builds upon the cycle-consistent adversarial network (CycleGAN) framework by introducing a paired complementary temporal method specifically designed for radar-based precipitation nowcasting.

While GAN-based approaches excel at generating high-fidelity, visually realistic precipitation patterns through its adversarial framework, they remain difficult to optimize due to persistent training instabilities and mode collapse. Furthermore, we also noticed that GANs have been predominantly restricted to domain-specific or regional applications; their scalability and effectiveness for global-scale weather prediction remain significant open questions in the field.

Diffusion-based models. `PrecipDiff` [Dai and Ushijima-Mwesigwa, 2025] leverages the power of diffusion models and residual learning to address the predictive accuracy, bias, and low spatial resolution. `DiffCast` [Yu *et al.*, 2024] models the precipitation process from two perspectives: the deterministic component accounts for predicting a global motion trend by a coarse forecast, while the stochastic component aims to learn local stochastic variations with the residual mechanism. `CasCast` [Gong *et al.*, 2024] develops a cascaded framework consisting of a deterministic predictive model to output blurry predictions, and a probabilistic diffusion model with inputs as both past observations and deterministic predictions beforehand. Because the deterministic predictions are the future frames, such frame-wise guidance in the diffusion model can provide a frame-to-frame correspondence between blurry predictions and latent vectors, resulting in a better generation of small-scale patterns. However, directly applying diffusion models might generate physically implausible predictions. To tackle these limitations, `Prediff` [Gao *et al.*, 2023] proposes a conditional latent diffusion model for probabilistic forecasts and then aligns forecasts with domain-specific physical constraints. This is achieved by estimating the deviation from imposed constraints at each denoising step and adjusting the transition distribution accordingly.

3.3 Pre-training & Fine-tuning-based Weather/Climate Foundation Models

Foundation Models (FMs) have garnered significant research interest due to their powerful prior knowledge acquired through pre-training on massive data and their remarkable adaptability to downstream tasks with fine-tuning strategies. The pre-training and fine-tuning process differentiates them from those models surveyed in Sec. 3.1 and 3.2. We review a few weather foundation models below.

`ClimaX` [Nguyen *et al.*, 2023] is a versatile and generalizable deep-learning model developed for weather and climate science. It is trained on heterogeneous datasets encompassing diverse variables, spatiotemporal coverage, and physical principles with CMIP6 datasets and it can be fine-tuned for a wide range of weather and climate applications, including those involving atmospheric variables and spatiotemporal scales not encountered during pre-training. `W-MAE` [Man *et al.*, 2023] is pre-trained with similar data, but using reconstruction tasks with the Masked Autoencoder model. The pre-trained model can be fine-tuned for various tasks, e.g., multi-variate prediction. `Aurora` [Bodnar *et al.*, 2025] is a large-scale foundation model pre-trained on over a million hours of diverse weather and climate data. Unlike the two foundation models above, `Aurora` can be fine-tuned in one of two ways: short-time fine-tuning (i.e., fine-tuning the entire architecture through one or two roll-out steps) and rollout fine-tuning for long-term multi-step predictions with low-rank adaptation (LoRA). `Prithvi WxC` [Schmude *et al.*, 2024] is a large-scale foundation model featuring 2.3 billion parameters and supporting 160 atmospheric variables. Its architecture utilizes a scalable 2D Vision Transformer (ViT) that employs flexible token and window sizes. During pre-training, a Masked Autoencoder (MAE) framework is used to mask varying ratios of

Methods	Resolution	Training data	Compute resources	Test data	Inference time	T850 $[K]$	Z500 $[m^2/s^2]$	U10m $[m/s]$	V10m $[m/s]$
Physics-based Models									
IFS HRES [ECMWF,]	0.1°			ERA5 2020	∼52 mins	2.23	411.07	3.05	3.17
IFS ENS [ECMWF,]	0.2°			ERA5 2020	–	1.9	360.85	2.51	3.61
Deterministic Predictive Models									
Pangu-Weather [Bi *et al.*, 2023]	0.25°	ERA5 1979-2017	16 days; 192 V100 GPUs	ERA5 2020	∼secs; a GPU	2.11	394.96	2.84	3.11
GraphCast [Lam *et al.*, 2022]	0.25°	ERA5 1979-2019	4 weeks; 32 TPU v4	ERA5 2020	∼min; a TPU	1.98	375.62	2.71	2.82
FuXi [Chen *et al.*, 2023b]	0.25°	ERA5 1979-2015	8 days; 8 A100 GPUs	ERA5 2020	∼secs; a GPU	1.84	352.74	2.43	2.54
Fengwu [Chen *et al.*, 2023a]	0.25°	ERA5 1979-2017	17 days; 32 A100 GPUs	ERA5 2020	∼secs; a GPU	1.97	365.62	–	–
Stormer [Nguyen *et al.*, 2024a]	0.25°	ERA5 1979-2017	8 days; 8 A100 GPUs	ERA5 2020	∼secs; a GPU	1.92	373.75	2.55	2.68
HEAL-ViT [Ramavajjala, 2024]	0.25°	ERA5 1979-2017	8 days; 8 A100 GPUs	ERA5 2020	∼secs; a GPU	1.99	369.99	2.72	2.98
GnnWeather [Keisler, 2022]	1°	ERA5 35 years	5.5 days; 1 A100 GPU	ERA5 2020	∼secs; a GPU	2.14	403.84	–	–
ArchesWeather[Couairon *et al.*, 2024]	1.5°	ERA5 1979-2018	9 days; 1 V100 GPU	ERA5 2020	∼secs; a GPU	1.98	381.05	2.66	2.79
NeuralGCM 0.7 [Kochkov *et al.*, 2024]	0.7°	ERA5 1979-2017	3 weeks; 256 TPUs v5	ERA5 2020	∼min; a TPU	1.98	363.88	–	–
NeuralGCM ENS [Kochkov *et al.*, 2024]	1.4°	ERA5 1979-2017	10 days; 128 TPUs v5	ERA5 2020	∼min; a TPU	1.82	345.86	–	–
Probabilistic Generative Models									
GenCast [Price *et al.*, 2025]	0.25°	ERA5 1979-2018	5 days; 32 TPUs v5	13	8 mins; a TPU	1.78	342.04	2.38	2.49
ArchesWeatherGen[Couairon *et al.*, 2024]	1.5°	ERA5 1979-2018	45 days; 1 V100 GPU	13	–	1.81	351.64	2.41	2.53
Foundation Models with Pre-training and Fine-tuning									
Aurora [Bodnar *et al.*, 2025]	0.1°; 0.25°	ERA5, CMIP6, ...	2.5 weeks; 32 A100 GPUs	HRES-T0 2022	–	∼1.85	∼350.22	∼2.67	–
Prithvi WxC [Schmude *et al.*, 2024]	0.5°	MERRA-2 1980-2019	–; 64 A100 GPUs	MERRA-2 2020-2023	–	∼2.25	–	∼3.15	–

Table 1: Comparison of approaches for **global** weather prediction. The performance RMSE scores (in the right four columns) for representative atmospheric variables (T850, Z500, U10m, and V10m) are reported at the lead time of 6 days (except Prithvi WxC at the lead time of 5 days). These scores are either from the WeatherBench scoreboard or the original paper. "Δx" represents the horizontal resolution.

these tokens and windows, enabling the model to effectively capture both localized regional dynamics and broader global dependencies. This versatile backbone can be fine-tuned for a diverse range of downstream applications, including nowcasting, medium-range prediction, and statistical downscaling. Recently, `AtmosArena` [Nguyen *et al.*, 2024b] benchmarks foundation models for atmospheric sciences across various atmospheric variables and lead times to predict.

3.4 Quantitative Comparison and Discussion

Table 1 presents a quantitative comparison of a part of the aforementioned models. Firstly, probabilistic generative and foundation models are emerging as promising directions, although they remain in the early stages of development. Notably, `GenCast` has achieved state-of-the-art performance, surpassing many deterministic models while producing probabilistic forecasts. However, existing probabilistic models are primarily based on diffusion models, resulting in significantly longer inference times than deterministic alternatives. Secondly, foundation models are distinguished by pre-training on diverse and large-scale datasets (e.g., ERA5, CMIP6, HRTS-T0), setting them apart from previous methods that rely on a single dataset. This diversity enhances their generalizability and benefits the subsequent fine-tuning phase, allowing them to adapt effectively to new tasks and domains. Thirdly, while ML-based models typically involve longer training durations and require substantial computational resources, they offer substantial speedups at inference time compared to traditional physics-based models. For instance, the inference process of models `Pangu-Weather` or `GraphCast` takes seconds to minutes on modern GPUs or TPUs, in contrast to 52 minutes required by IFS HRES. Fourthly, we observe a nuanced relationship between spatial resolution (Δx) and forecast skill. While high-resolution models such as `Pangu-Weather` and `GraphCast` (0.25°) generally lead the benchmarks, certain low-resolution architectures, notably `GnnWeather` and `ArchesWeather` (1.5°), maintain competitive performance for Z500. This performance suggests that architectural innovations can effectively compensate for coarser spatial grids without a proportional loss in predictive accuracy.

3.5 Representative Applications

We provide an overview of the available datasets in Table 2. Each data set is presented by domain, name, coverage, collection method, spatial and temporal resolution, and time span.

4 Challenges and Future Directions

We introduce primary challenges and suggest novel research opportunities from the perspectives of *data* (§4.1-4.2), *model* (§4.3-4.5), and *operational deployment* (§4.6).

4.1 Data Processing and Management

The volume of weather and climate data is increasing rapidly. European Centre for Medium-Range Weather Forecasts (ECMWF) archives contain hundreds of PB of data increasing daily. Variational Autoencoder (VAE) approaches have emerged as powerful tools for data compression, converting the high-dimensional data from the original space to a lower latent space. `WeatherVAE` [Liu *et al.*, 2024b] reduces the data size from 8.61 TB to a compact 204 GB, and `VAEformer` [Han *et al.*, 2024] compresses the ERA5 dataset (226 TB) into a CRA5 dataset (0.7 TB). More importantly, they demonstrate that downstream experiments of global weather prediction models trained on the compact CRA5 dataset achieve accuracy comparable to the models trained on the original dataset. This approach significantly reduces storage requirements for massive weather datasets.

4.2 Weather Knowledge Distillation

While these large models [Price *et al.*, 2025; Bodnar *et al.*, 2025] yield impressive accuracies, they incur substantial training time and memory overhead, making them challenging to fine-tune or train from scratch. Knowledge distillation is a well-established technique in which a smaller "student" model is trained to replicate the outputs or internal representations of a larger "teacher" model. Such an approach significantly reduces computational resource needs in terms of processing and memory during inference while maintaining competitive predictive performance. Given the substantial computational demands of large-scale weather prediction models, applying distillation to this domain presents a

Domain	Dataset	Coverage	Collect	Spatial	Temporal	Time Span
General Weather	WeatherBench	Global	Reanalysis	$1.40625°, 2.8125°, 5.625°$	6 hours	1979-2018
	WeatherBench 2	Global	Reanalysis	$0.25°$	6 hours	1979-2020
	Weather5K	Global	Observation	-	1 hour	2014-2023
Precipitation	SEVIR	Region in U.S.	Radar&Satellite	1 km$\times$1 km	5 mins	2017-2019
	Meteonet	France	Radar&Satellite	1 km	5-15 mins	2016-2018
	HKO-7	Region in Hong Kong	Radar	1 km$\times$1 km	6 mins	2009-2015
Wind	GlobalWindTemp	Global	Observation	-	1 hour	2019-2010
	DigitalTyphoon	W.N. Pacific basin	Satellite	5 km	1 hour	1978-2022
	TropicalCyclone	Global	CAM5 simulation	25 km	3 hours	1979-2005
Air Quality	UrbanAir	Regional, China	Observation	-	1 hour	2014-2015
	KnowAir	Regional, China	Observation	-	3 hours	2015-2018
	BeijingAir	Regional, China	Observation	-	1 hour	2010-2014
SST	TropicalOcean	Pacific Ocean	Observation	5°S-5°N, 120°E-80°W	Monthly	1982–2017
	CMIP6	Tropical Pacific	Simulation	$2° \times 0.5°$	Monthly	1850–2014
	NOAA/CIRE	Global	Reanalysis	$2° \times 2°$	6 hours	1850–2015
	ENSO	Tropical Pacific	NOAA, NCEI, NCAR	-	Monthly	1950–2023
Flood	DEM	Carlisle, UK	Observation	5 m	1 hour	2005-2015
	AustraliaFlood	Australia	Observation	-	Daily	1900-2018
	FloridaFlood	South Florida, US	Observation	-	1 hour	2010-2020
	QijiangRiverBasin	QijiangRiver, China	Observation	-	1 hour	1979-2020

Table 2: Summary of representative data sets on weather prediction. CAM5: Community Atmospheric Model v5. We include more related work in our curated GitHub repository (https://github.com/JimengShi/DL-Foundation-Models-Weather), due to the limited space.

promising avenue for exploration. Researchers from NVIDIA published a related work [Martin *et al.*, 2025] in this context.

4.3 Probabilistic Models for Weather Uncertainty

Given the chaotic nature of the atmosphere, quantifying atmospheric uncertainty is vital for informed decision-making. Approaches such as initial conditions perturbation and Monte Carlo dropout often fail to simultaneously account for both *aleatoric* uncertainty from the inherent randomness in weather data and *epistemic* uncertainty from models themselves due to limited knowledge. In contrast, the probabilistic nature of diffusion models allows them to learn the full data distribution. By utilizing stochastic sampling and noise conditioning, they inherently capture both data randomness and model uncertainty. Given the scarcity of existing literature and the precious value of diffusion models for weather modeling, we believe it deserves further investigation.

4.4 Weather Prediction and Understanding

Most existing approaches focus on predicting *numerical* meteorological variables (e.g., temperature). Despite superior predictive performance, they often isolate the understanding or interpretation in natural languages (e.g., a report). Therefore, further research on multi-modal models is needed for weather prediction and understanding. Although pioneer studies, such as `Omni-Weather` [Zhou *et al.*, 2025] and `RadarQA` [He *et al.*, 2025], have recently begun exploring this domain, the integration of multimodal reasoning into weather prediction is still in its infant stage.

4.5 Agentic Weather System

Subsection 4.5 discusses weather understanding in natural languages. However, it is quite challenging to build an interactive system that has strong executive capabilities to handle the complex queries that demand multiple databases, models, and/or tools. As AI agents emerge, agentic frameworks can integrate and interact with multiple external resources and tools, which offers a promising solution. `ZephyrusWorld` [Varambally *et al.*, 2025] proposes a multi-turn LLM-based weather agent that iteratively analyzes weather datasets, observes results, and refines its approach through conversational feedback loops. This track has become popular in language-based tasks; it is vital to explore it further in weather science.

4.6 Operational Deployment

Despite their superior performance, integrating AI models into operational pipelines introduces critical challenges. The opacity of these "black-box" architectures hinders model verification, making it difficult for forecasters to contest predictions and potentially undermining the trust required for high-stakes decisions. To mitigate these risks, systems must prioritize robust uncertainty quantification and interpretable outputs within a human-in-the-loop framework, allowing experts to validate or contextualize model guidance. Furthermore, establishing accountability through data provenance, version control, and traceable prediction logs is essential for ensuring reproducibility and transparency in public safety sectors.

5 Conclusions

This work provides a comprehensive and up-to-date survey of data-driven models for weather prediction. We propose a novel taxonomy categorized by model training paradigms, facilitating a rigorous in-depth review and comparative analysis of existing methods. Beyond architectural discussion, we summarize available datasets, open-source codebases, and diverse real-world applications in a GitHub repository. We also analyze emerging challenges and propose critical research directions for advancing AI-enabled weather prediction.

Acknowledgments

This work is supported by the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE), funded by the U.S. National Science Foundation (NSF) under award number 2118329. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily represent the views of NSF.

References

[Abbass *et al.*, 2022] Kashif Abbass, M Zeeshan Qasim, and et al. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, 29(28), 2022.

[Alexe *et al.*, 2024] Mihai Alexe, Eulalie Boucher, et al. Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations, 2024.

[Andrychowicz *et al.*, 2023] Marcin Andrychowicz, Lasse Espeholt, et al. Deep learning for day forecasts from sparse observations, 2023.

[Bai *et al.*, 2022] Cong Bai, Feng Sun, Jinglin Zhang, et al. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[Bi *et al.*, 2023] Kaifeng Bi, Lingxi Xie, et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023.

[Bodnar *et al.*, 2025] Cristian Bodnar, Wessel P Bruinsma, and Ana others Lucic. A foundation model for the earth system. *Nature*, pages 1–8, 2025.

[Chen *et al.*, 2023a] Kanget Chen, Tao Han, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead, 2023.

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

[Chen *et al.*, 2024] Yutong Chen, Ya Wang, Gang Huang, and Qun Tian. Coupling physical factors for precipitation forecast in China with graph neural network. *Geophysical Research Letters*, 51(2), 2024.

[Choi *et al.*, 2023] Jaeho Choi, Yura Kim, Kwang-Ho Kim, et al. PCT-CycleGAN: Paired complementary temporal cycle-consistent adversarial networks for radar-based precipitation nowcasting. In *CIKM*, 2023.

[Couairon *et al.*, 2024] Guillaume Couairon, Renu Singh, et al. Archesweather & archesweathergen: a deterministic and generative model for efficient weather forecast, 2024.

[Dai and Ushijima-Mwesigwa, 2025] Ting-Yu Dai and Hayato Ushijima-Mwesigwa. Precipdiff: Leveraging image diffusion models to enhance satellite-based precipitation observations. In *AAAI*, 2025.

[Das *et al.*, 2024] Puja Das, August Posch, et al. Hybrid physics-AI outperforms numerical weather prediction for extreme precipitation nowcasting, 2024.

[ECMWF,] ECMWF. https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts.

[Espeholt *et al.*, 2022] Lasse Espeholt, Shreya Agrawal, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):1–10, 2022.

[Gao *et al.*, 2022] Zhihan Gao, Xingjian Shi, et al. Earthformer: Exploring space-time transformers for earth system forecasting. *In NeurIPS*, 2022.

[Gao *et al.*, 2023] Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, et al. Prediff: Precipitation nowcasting with latent diffusion models. In *NeurIPS*, 2023.

[Gong *et al.*, 2024] Junchao Gong, Lei Bai, et al. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. In *ICML*, 2024.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Guibas *et al.*, 2021] John Guibas, Morteza Mardani, et al. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

[Han *et al.*, 2024] Tao Han, Zhenghao Chen, et al. CRA5: Extreme compression of ERA5 for portable global climate and weather research via an efficient variational transformer. *arXiv preprint arXiv:2405.03376*, 2024.

[Hatanpää *et al.*, 2025] Väinö Hatanpää, Eugene Ku, Jason Stock, et al. Aeris: Argonne earth systems model for reliable and skillful predictions. In *ICHPCNSA*, 2025.

[He *et al.*, 2025] Xuming He, Zhiyuan You, et al. Radarqa: Multi-modal quality analysis of weather radar forecasts, 2025.

[Keisler, 2022] Ryan Keisler. Forecasting global weather with graph neural networks, 2022.

[Kochkov *et al.*, 2024] Dmitrii Kochkov, Janni Yuval, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

[Lam *et al.*, 2022] Remi Lam, Alvaro Sanchez, et al. GraphCast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[Lang *et al.*, 2024] Simon Lang, Mihai Alexe, et al. AIFS-ECMWF's data-driven forecasting system, 2024.

[Liu *et al.*, 2024a] Peiyuan Liu, Tian Zhou, et al. Mitigating time discretization challenges with weatherODE: A sandwich physics-driven neural ODE for weather forecasting. *arXiv preprint arXiv:2410.06560*, 2024.

[Liu *et al.*, 2024b] Qian Liu, Bing Gong, et al. Compressing high-resolution data through latent representation encoding for downscaling large-scale AI weather forecast model. *arXiv preprint arXiv:2410.09109*, 2024.

[Liu *et al.*, 2024c] Tengyuan Liu, Lei Liu, et al. WPMamba: Enhanced wind power forecasting model based on Mamba with weather forecast data. In *ICPRE*, 2024.

[Liu *et al.*, 2024d] Yong Liu, Tengge Hu, et al. iTransformer: Inverted transformers are effective for time series forecasting. In *ICLR*, 2024.

[Liu *et al.*, 2024e] Zili Liu, Hao Chen, et al. MambaDS: Near-surface meteorological field downscaling with topography constrained selective state space modeling. *IEEE TGRG*, 2024.

[Luo *et al.*, 2022] Chuyao Luo, Xutao Li, et al. Experimental study on generative adversarial network for precipitation nowcasting. *IEEE TGRS*, 60:1–20, 2022.

[Ma *et al.*, 2023a] Minbo Ma, Peng Xie, et al. HiSTGNN: Hierarchical spatio-temporal graph neural network for weather forecasting. *Information Sciences*, 648, 2023.

[Ma *et al.*, 2023b] Zhifeng Ma, Hao Zhang, and Jie Liu. MM-RNN: A multimodal RNN for precipitation nowcasting. *IEEE TGRS*, 2023.

[Man *et al.*, 2023] Xin Man, Chenghong Zhang, et al. W-MAE: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting, 2023.

[Martin *et al.*, 2025] Scott A Martin, Noah Brenowitz, Dale Durran, and Michael Pritchard. Long-range distillation: Distilling 10,000 years of simulated climate into long timestep ai weather models. *arXiv preprint arXiv:2512.22814*, 2025.

[Nguyen *et al.*, 2023] Tung Nguyen, Johannes Brandstetter, et al. ClimaX: A foundation model for weather and climate. In *ICML*, 2023.

[Nguyen *et al.*, 2024a] Tung Nguyen, Rohan Shah, et al. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. 2024.

[Nguyen *et al.*, 2024b] Tung Nguyen, Prateik Sinha, et al. Atmosarena: Benchmarking foundation models for atmospheric sciences. In *NeurIPS Workshop*, 2024.

[Pathak *et al.*, 2022] Jaideep Pathak, Shashank Subramanian, et al. A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022.

[Price *et al.*, 2025] Ilan Price, Alvaro Sanchez-Gonzalez, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

[Qin *et al.*, 2024] Haoyu Qin, Yungang Chen, et al. MetMamba: Regional weather forecasting with spatial-temporal mamba model, 2024.

[Ramavajjala, 2024] Vivek Ramavajjala. HEAL-ViT: Vision transformers on a spherical mesh for medium-range weather forecasting, 2024.

[Rasp *et al.*, 2023] Stephan Rasp, Stephan Hoyer, Alexander Merose, et al. WeatherBench 2: A benchmark for the next generation of data-driven global weather models, 2023.

[Ravuri *et al.*, 2021] Suman Ravuri, Karel Lenc, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878), 2021.

[Saleem *et al.*, 2024] Hira Saleem, Flora Salim, et al. Conformer: Embedding continuous attention in vision transformer for weather forecasting. In *ICLR*, 2024.

[Schmude *et al.*, 2024] Johannes Schmude, Sujit Roy, et al. Prithvi WxC: Foundation model for weather and climate. 2024.

[Shi *et al.*, 2025] Jimeng Shi, Bowen Jin, et al. Codicast: Conditional diffusion model for global weather forecasting with uncertainty quantification. In *IJCAI*, 2025.

[Tang *et al.*, 2024] Yujin Tang, Peijie Dong, et al. VMRNN: Integrating Vision Mamba and LSTM for efficient and accurate spatiotemporal forecasting. In *CVPR*, 2024.

[Varambally *et al.*, 2025] Sumanth Varambally, Marshall Fisher, et al. Zephyrus: An agentic framework for weather science, 2025.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. In *NeurIPS*, 2017.

[Verma *et al.*, 2024] Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In *ICLR*, 2024.

[Wang *et al.*, 2018] Yunbo Wang, Zhifeng Gao, et al. Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *ICML*, 2018.

[Wang *et al.*, 2022] Yunbo Wang, Haixu Wu, et al. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE TPAMI*, 45(2), 2022.

[Wu and others, 2024] Binqing Wu et al. WeatherGNN: Exploiting meteo-and spatial-dependencies for local numerical weather prediction bias-correction. In *IJCAI*, 2024.

[Xiao *et al.*, 2023] Yi Xiao, Lei Bai, Wei Xue, et al. Fengwu-4DVar: Coupling the data-driven weather forecasting model with 4D variational assimilation, 2023.

[Xu *et al.*, 2024] Luwen Xu, Jiwei Qin, Dezhi Sun, et al. PFformer: A time-series forecasting model for short-term precipitation forecasting. *IEEE Access*, 2024.

[Yang *et al.*, 2024] Qidong Yang, Jonathan Giezendanner, et al. Multi-modal graph neural networks for localized off-grid weather forecasting. 2024.

[Yu *et al.*, 2024] Demin Yu, Xutao Li, Yunming Ye, et al. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *CVPR*, 2024.

[Yuan *et al.*, 2025] Shijin Yuan, Guansong Wang, et al. Tianxing: A linear complexity transformer model with explicit attention decay for global weather forecasting. *Advances in Atmospheric Sciences*, 42(1):9–25, 2025.

[Yuval *et al.*, 2026] Janni Yuval, Ian Langmore, et al. Neural general circulation models for modeling precipitation. *Science Advances*, 12(2), 2026.

[Zhang *et al.*, 2023] Yuchen Zhang, Mingsheng Long, et al. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619, 2023.

[Zhao *et al.*, 2024] Pengcheng Zhao, Jiang Bian, et al. OMG-HD: A high-resolution AI weather model for end-to-end forecasts from observations, 2024.

[Zhou *et al.*, 2025] Zhiwang Zhou, Yuandong Pu, et al. Omni-weather: Unified multimodal foundation model for weather generation and understanding, 2025.