DIFFUSION PRECIPITATION GENERATION WITH MEDIUM-RANGE WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate daily precipitation forecasts are important for weather-sensitive decisions. Although recent Weather Foundation Models (WFMs) have achieved significant progress, they are trained on reanalysis data, which exhibit regional and intensity-dependent biases. Since retraining WFMs on observation-based targets is costly, we study *Dynamics–Precipitation Decoupling*, a plug-and-play framework that uses a separately trained generator to map non-precipitation forecasts from a frozen weather model to observation-aligned daily precipitation. Its key challenges are to capture heterogeneous rainfall distributions and align daily accumulations with atmospheric signals that evolve across space and time. To address them, we train a diffusion generator conditional on atmospheric states and satellite-gauge estimates of precipitation. A multi-scale recurrent encoder summarizes atmospheric evolution, while cross-attention aligns atmospheric and precipitation features during denoising. Extensive evaluation of various backbones, including Pangu-Weather, GraphCast, and HRES, on global and regional forecasting shows improved daily precipitation accuracy and event detection, with better representation of localized accumulation patterns. Global ensemble precipitation forecasting results demonstrate that our decoupling models outperform ENS and training-from-scratch GenCast. Remarkably, a heavy-rainfall case study using 2,288 stations in China further supports improved daily accumulation patterns and event detection against ground observations. Code is available at https://anonymous.4open.science/r/ICLR2027-code-9B4C.

1 INTRODUCTION

Accurate daily precipitation forecasting is a core task in weather prediction, playing a pivotal role in disaster preparedness, agricultural production, transportation management, and water-resource planning (Hou et al., 2014; Vogel et al., 2019). Traditional approaches mainly rely on Numerical Weather Prediction (NWP), which solves complex atmospheric dynamical and thermodynamic equations to simulate atmospheric evolution, or statistical methods, which mine relationships between historical precipitation and meteorological factors such as relative humidity (Bauer et al., 2015; Palmer, 2017; Sun et al., 2018; Wilks, 2011; Benziane & MB, 2024). However, NWP systems are computationally intensive and remain sensitive to uncertainties in initial conditions and physical parameterizations (Bauer et al., 2021). These limitations increasingly motivate the development of more data-driven forecasting approaches. Initially, and in most subsequent precipitation studies, models were trained on observational data, such as radar sequences, and focused on short-term forecasts over local regions, referred to as precipitation nowcasting (An et al., 2025; Leinonen et al., 2023; Gong et al., 2024). Recent development of Weather Foundation Models (WFMs) trained on ERA5 (Hersbach et al., 2020), including FourCastNet (Pathak et al., 2022) and GenCast (Price et al., 2025), has enabled the exploration of global medium-range precipitation forecasting.

Nevertheless, precipitation in ERA5 is produced through reanalysis model physics and data assimilation, and previous evaluations have documented **regional and intensity-dependent biases**, including systematic over- or underestimation and distorted spatiotemporal distributions (Xie et al., 2022; Jiao et al., 2021). Consequently, localized precipitation extremes are often smoothed or underrepresented despite accurate large-scale atmospheric forecasts. To obtain a more direct observation-aligned precipitation target, we use the **I**ntegrated **M**ulti-satellit**E** **R**etrievals for **G**PM (IMERG) satellite-gauge product (Pradhan et al., 2022). A straightforward solution would be to replace the precipitation target

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

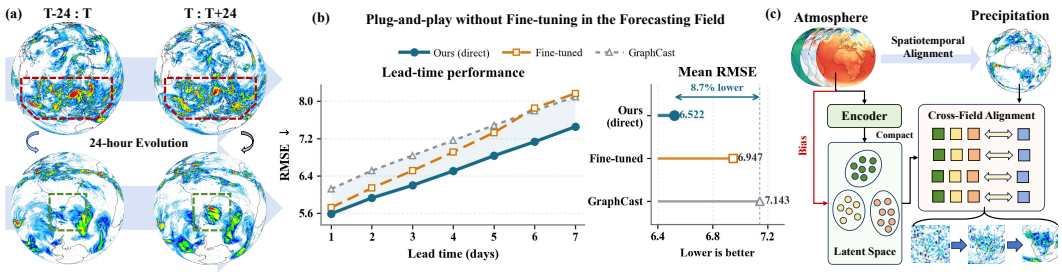


Figure 1: Motivation and overview. (a) Consecutive daily precipitation fields exhibit substantial variability. (b) Directly plugging the generator into forecast models outperforms forecast-field fine-tuning. (c) The encoder compresses atmospheric variables into spatiotemporal features, followed by cross-field alignment with precipitation to guide generation. Together, these observations motivate our Dynamics–Precipitation Decoupling paradigm and direct plug-in design for precipitation generation.

and retrain WFMs; however, this is impractical because of the prohibitive training cost. For instance, training GraphCast required about four weeks on 32 Cloud TPU v4 devices (Lam et al., 2023).

Therefore, we study *Dynamics–Precipitation Decoupling*: a separately trained generator maps non-precipitation forecasts from a frozen backbone to observation-aligned daily precipitation. This requires addressing two challenges:

- Daily precipitation exhibits heterogeneous spatial patterns and skewed accumulation distributions (Liu et al., 2025; Gao et al., 2022), as shown in Figure 1(a). Under conditional uncertainty, deterministic models trained with pointwise regression can favor conditional averages, smoothing local accumulation maxima and rainband structures (Ravuri et al., 2021).
- Atmospheric conditioning poses a nonlocal spatiotemporal alignment challenge. Daily precipitation reflects the cumulative outcome of evolving atmospheric processes, whereas key predictors such as moisture transport and vertical motion may occur earlier and over surrounding or upstream regions (Trenberth et al., 2003). As a result, precipitation at a given location cannot be explained by atmospheric conditions at the same grid point and time alone.

Recent studies have shown the potential of generative models to preserve localized structures and sharpen precipitation boundaries in radar nowcasting and regional downscaling (Gao et al., 2023; Yu et al., 2024; Gong et al., 2024; Liu et al., 2025). In this work, we propose a plug-and-play conditional diffusion generator addresses this challenge through two components, as shown in Figure 1(c), which consists of: (1) a multi-scale recurrent encoder captures atmospheric evolution in compact, daily precipitation features; (2) cross-attention aligns atmospheric and precipitation features across scales to guide denoising. The generator learns from paired ERA5 and IMERG precipitation, using only non-precipitation states as conditioning inputs. During inference, its atmospheric conditions are provided by a WFM or NWP system, and a noisy precipitation field is iteratively refined through a reverse-diffusion process under the guidance of encoded atmospheric features. Preliminary experiments in Figure 1(b) show that such a model, without backbone-specific retraining, outperforms both the backbone's native precipitation forecasts and forecast-field fine-tuning. Extensive experiments cover three complementary levels: global forecasting at $1.5°$, regional forecasting at $0.25°$, and a heavy-rainfall case study using observations from 2,288 stations in China. Across data-driven and numerical backbones, we evaluate deterministic skill, probabilistic accuracy, input sensitivity, and module efficiency. Station observations further support improved daily accumulation patterns and event detection beyond IMERG-based evaluation. Our main contributions are summarized as follows:

- We highlight **Dynamics–Precipitation Decoupling**, separating atmospheric evolution from observation-supervised precipitation generation and enabling IMERG-aligned learning without retraining a complete weather forecasting system.
- We develop a forecast-conditioned diffusion architecture that combines a multi-scale atmospheric encoder with an attention-conditioned denoiser to model 24-hour precipitation distributions.
- We reuse one ERA5-trained generator with aligned forecasts from frozen Pangu-Weather, GraphCast, and HRES, outperforming available native precipitation forecasts without training a full weather model from scratch.
- We provide complementary global, regional, and station-level evaluations of forecast skill and input sensitivity. Validation against 2,288 stations shows improved local rainfall intensity patterns and event detection during the evaluated heavy-rainfall event.

2 Related Work

Precipitation forecasting with machine learning Existing learning-based precipitation forecasting spans two main settings (Gao et al., 2021; Ashok & Pekkat, 2022; Ayzel et al., 2020). Radar-based nowcasting extrapolates recent observations over regional domains, with representative methods including ConvLSTM (Shi et al., 2015), Rainformer (Bai et al., 2022), and NowcastNet (Zhang et al., 2023). These methods capture local precipitation evolution but generally target lead times of only a few hours. WFMs such as FourCastNet (Pathak et al., 2022) and GenCast (Price et al., 2025) extend data-driven forecasting to global domains and medium-range lead times, while FuXi-S2S and ClimaX address subseasonal and climate-oriented tasks (Chen et al., 2024; Nguyen et al., 2023; Watson-Parris et al., 2022). Their precipitation products nevertheless remain tied to system-specific supervision, and documented biases in ERA5 precipitation can limit agreement with satellite–gauge observations (Xie et al., 2022; Jiao et al., 2021). The most related study is NeuralGCM (Yuval et al., 2026), which also utilizes IMERG supervision at $2.8°$. However, our work addresses a different setting: rather than integrating precipitation prediction into the forecasting backbone, we decouple precipitation generation from atmospheric forecasting. A separately trained generator maps non-precipitation forecast states to observation-aligned 24-hour precipitation, allowing the same generator to be reused across aligned WFM and NWP backbones without backbone-specific retraining. This design targets medium-range precipitation forecasting at finer spatial resolutions of $1.5°$ and $0.25°$, with particular emphasis on localized rainfall structures.

Diffusion models for weather forecasting Diffusion-based weather studies cover full-state forecasting, regional atmospheric emulation, and precipitation nowcasting (Bauer et al., 2015; Tyralis & Papacharalampous, 2024). GenCast (Price et al., 2025) generates probabilistic global atmospheric trajectories, while Continuous Ensemble (Andrae et al., 2025) produces temporally consistent ensembles at flexible lead times. StormCast (Pathak et al., 2026) and HRRRCast (Abdi et al., 2025) emulate convection-allowing regional evolution, whereas PreDiff and DiffCast (Gao et al., 2023; Yu et al., 2024) generate future radar sequences from recent precipitation observations. These methods therefore evolve complete atmospheric states, emulate a particular regional system, or extrapolate precipitation history. Our framework instead retains an existing WFM or NWP backbone and performs cross-variable generation, using sequences of non-precipitation forecast states to condition an independently IMERG-supervised diffusion generator.

Precipitation post-processing and downscaling Precipitation post-processing commonly calibrates or corrects existing precipitation forecasts using observations (Hewson & Pillosu, 2021; Liu et al., 2023; Tian et al., 2024; Zhang et al., 2025). Same-variable downscaling reconstructs finer precipitation fields from coarse satellite or model estimates (Zhu et al., 2023; 2025; Ibrahim et al., 2026). CorrDiff (Mardani et al., 2025) extends this setting to multivariate downscaling and additional field synthesis through regression and residual diffusion. These approaches generally transform or refine an existing precipitation field while we generate observation-aligned precipitation directly from non-precipitation atmospheric states, without requiring a precipitation forecast as input. The task is therefore neither precipitation correction nor spatial super-resolution, but observation-supervised cross-variable generation on the conditioning grid.

3 Background

Problem definition. We study the weather forecasting (Lam et al., 2023; Kochkov et al., 2024; Price et al., 2025) of V weather parameters at the latitude-longitude grid $\boldsymbol{G} \in \mathbb{R}^{H\times W\times 2}$, where H and W are the height and width of the grid that depend on the resolution of latitude and longitude, and $\boldsymbol{G}_{h,w,:} = (\lambda_h, \phi_w) \in \Omega = [-90°, 90°] \times [-180°, 180°]$. At hour t, the global non-precipitation atmospheric state is represented by $\boldsymbol{X}_t \in \mathbb{R}^{H\times W\times V}$. Here, $V = V_A C + V_S$, where V_A, V_S, and C denote the numbers of upper-air variables, surface variables, and pressure levels, respectively. Given the initial weather conditions $(\boldsymbol{G}, \boldsymbol{X}_0)$, our goal is to learn a neural network to forecast 24-hour accumulated precipitation at future time horizons, as shown in the following:

$$(\hat{\boldsymbol{Y}}_{24}, \hat{\boldsymbol{Y}}_{48}, \cdots, \hat{\boldsymbol{Y}}_T) = f_\Theta(\boldsymbol{X}_0), \tag{1}$$

where Θ denotes the parameters of the neural network. Each prediction $\hat{\boldsymbol{Y}}_t \in \mathbb{R}^{H\times W}$ represents the accumulated precipitation over the preceding 24-hour window $[t-24, t]$ with $t \in \{24, 48, \cdots, T\}$. In this work, we consider forecasts up to 7 days ahead ($T = 168$).

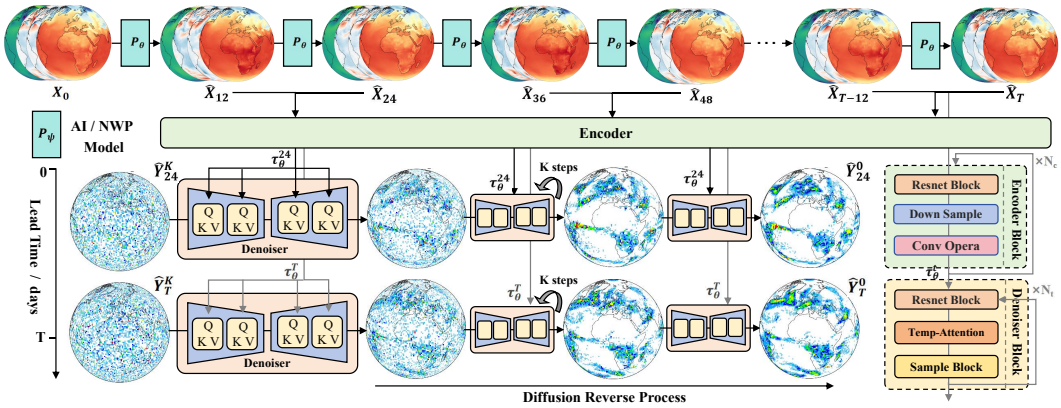


Figure 2: Overview of our two-step framework. The workflow cycles as follows: (1) An AI/NWP model produces two 12h autoregressive forecasts($\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t$), which the Encoder converts into the features. (2) Guided by these features, the Denoiser generates a 24h accumulated precipitation forecast $\hat{\boldsymbol{Y}}_t$ via a K-step reverse diffusion process. The details are shown in the bottom right.

Diffusion Model Diffusion models (Ho et al., 2020) are probabilistic frameworks designed to learn the target distribution $p(\boldsymbol{Y}_t)$ by reversing a fixed Markov Chain of length K. In our work, the reverse process transforms Gaussian noise into a precipitation field. Specifically, the diffusion model consists of two processes: a forward noising process and a reverse denoising process.

In the forward process, Gaussian noise is progressively injected into ground-truth precipitation fields $\boldsymbol{Y}_t^0 = \boldsymbol{Y}_t$ and the k-th diffusion state $\boldsymbol{Y}_t^k$ is defined as:

$$q(\boldsymbol{Y}_t^k|\boldsymbol{Y}_t^0) = \mathcal{N}(\boldsymbol{Y}_t^k; \sqrt{\overline{\alpha}_k}\boldsymbol{Y}_t^0, (1-\overline{\alpha}_k)I), \tag{2}$$

where $\overline{\alpha}_k = \prod_{j=1}^{k} \alpha_j$, $\alpha_j = 1 - \beta_j$, and $\beta_j \in (0,1)$ is predefined by a variance schedule. Regarding the reverse process, the target precipitation field is reconstructed via a Markov chain characterized by parameterized Gaussian transitions:

$$p_\theta(\boldsymbol{Y}_t^{k-1}|\boldsymbol{Y}_t^k) = \mathcal{N}(\boldsymbol{Y}_t^{k-1}; \mu_\theta(\boldsymbol{Y}_t^k, k), \sigma_k^2 I), \tag{3}$$

where μ_θ represents the posterior mean function. The reverse function $p_\theta(\boldsymbol{Y}_t^{k-1}|\boldsymbol{Y}_t^k)$ aims to remove the noise in the precipitation field. While this reverse transition is probabilistic, the most successful frameworks (Ho et al., 2020; Dhariwal & Nichol, 2021) leverage a variant of the variational lower bound, which mirrors denoising score-matching (Song et al., 2020). These models can be interpreted as a sequence of denoising methods $\epsilon_\theta(\boldsymbol{Y}_t^k, k)$, optimized to predict the noise $\epsilon \sim \mathcal{N}(0, I)$ in a degraded precipitation map $\boldsymbol{Y}_t$. Following the objective, the loss function is defined as:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{Y}_t^0, k, \epsilon} \left\| \epsilon - \epsilon_\theta(\boldsymbol{Y}_t^k, k) \right\|^2, \tag{4}$$

where the k-th denoising state $\boldsymbol{Y}_t^k = \sqrt{\overline{\alpha}_k}\boldsymbol{Y}_t^0 + \sqrt{1-\overline{\alpha}_k}\epsilon$.

4 Diffusion Precipitation Generation

4.1 Dynamics–Precipitation Decoupling

Instead of training a full weather prediction model from scratch, as illustrated in Figure 2, we decouple the task into: (1) obtaining future atmospheric forecasts; (2) generating precipitation fields via conditioned diffusion. This design allows our framework to reuse forecast atmospheric states from existing weather models to generate precipitation aligned with a new target product (e.g., IMERG) through a learnable mapping, without expensive end-to-end retraining across different systems.

In the first step, forecasts of non-precipitation atmospheric variables (e.g., temperature, wind, and humidity) are obtained from a base weather model, which can be either an NWP system (e.g., HRES) or a WFM (e.g., GraphCast and Pangu-Weather). The base model remains frozen and only provides conditioning fields, with no gradient propagated through it. As shown in Figure 1(b),

direct reuse of the ERA5-trained generator performs better than forecast-field fine-tuning, indicating that adapting to forecast states affected by accumulated lead-time errors can weaken the learned atmosphere–precipitation relationship. We therefore use ERA5 reanalysis states $\boldsymbol{X}_t$ during training and atmospheric forecasts $\boldsymbol{P}_\psi(\boldsymbol{X}_0, t)$ during prediction, where $\boldsymbol{P}_\psi(\cdot)$ is the base forecasting model with frozen parameters ψ. $\hat{\boldsymbol{X}}_0 = \boldsymbol{X}_0$ denotes the initial global weather state and $\hat{\boldsymbol{X}}_t$ represents the atmospheric variables at hour $t \in \{12, 24, \cdots, T-12, T\}$ in 12-hour intervals. For each 24-hour precipitation target, the two states at $t-12$ and t provide a compact description of the intraday atmospheric change. A denser sequence at 6-hour intervals is reserved for the temporal-density comparison in Appendix A.3 for more detailed analysis.

In the second step, this framework employs a conditional diffusion model to generate the 24-hour accumulated precipitation field conditioned on $(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)$. The diffusion process learns to reconstruct the precipitation distribution by gradually denoising from Gaussian noise. Formally, for any $t \in \{24, 48, \cdots, T\}$, we denote the target precipitation field as $\boldsymbol{Y}_t = \boldsymbol{Y}_t^0$, and let $\boldsymbol{Y}_t^1, \cdots, \boldsymbol{Y}_t^K$ represent the latent variables in the diffusion chain. Assuming a K-step denoising process, the conditional distribution factorizes as

$$p_\theta(\boldsymbol{Y}_t^{0:K} | \hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t) = p(\boldsymbol{Y}_t^K) \prod_{k=1}^{K} p_\theta(\boldsymbol{Y}_t^{k-1} | \boldsymbol{Y}_t^k, \hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t), \tag{5}$$

where $\boldsymbol{Y}_t^K \sim \mathcal{N}(0, I)$. This formulation estimates the conditional precipitation distribution at each forecast horizon under its corresponding atmospheric condition. In this formulation, the denoising process is no longer a blind reconstruction of noise but a constrained synthesis guided by the evolution of atmospheric states. We detail the framework design of the Encoder and the Conditional Denoiser that realizes this complex mapping f_Θ in the following section.

Training Based on the atmosphere-conditioning pairs, we then learn the diffusion precipitation generation via

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{Y}_t^0, \hat{\boldsymbol{X}}, k, \epsilon} \left\| \epsilon - \epsilon_\theta(\boldsymbol{Y}_t^k, k, \tau_\theta(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)) \right\|^2, \tag{6}$$

where k is sampled from $\{1, \ldots, K\}$, $\epsilon \sim \mathcal{N}(0, I)$, and both τ_θ and ϵ_θ are optimized via Eq. (6).

4.2 Precipitation Generation Architecture

As mentioned earlier, we integrate forecasts $\hat{\boldsymbol{X}}_t$ from NWP systems or WFMs $\boldsymbol{P}_\psi$ into the diffusion framework as conditioning information for generating the corresponding target precipitation field. Drawing inspiration from generative models (Rombach et al., 2022), diffusion models (DMs) are in principle capable of modeling conditional distributions of the form $p(\boldsymbol{Y}_t | \hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)$. This can be implemented with a conditional denoising method $\epsilon_\theta(\boldsymbol{Y}_t^k, k, \tau_\theta(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t))$. **However**, directly feeding raw multivariate atmospheric fields into the denoiser is memory-intensive due to their high dimensionality. Furthermore, sequential atmospheric states contain substantial spatiotemporal redundancy. To address this, we use a convolutional encoder to compress multi-level forcing into a latent representation, reducing computation while preserving precipitation-relevant signals.

Multi-Scale Atmospheric Encoder Following the previous studies (Trenberth et al., 2003; Bi et al., 2023; Espeholt et al., 2022), we propose an **Atmospheric Encoder** to effectively transform atmospheric forecasts $(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)$ into an intermediate latent representation: $\tau_\theta^t = \tau_\theta(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)$. As shown in Figure 2, each Encoder block consists of a ResNet block, a sampling operator, and a convolutional operator. First, let $l \in \{1, \ldots, L\}$ denote the index of the downsampling stage. The feature extraction at layer l is formulated as:

$$\boldsymbol{R}_t^l = \mathrm{Conv}\big(\sigma(\mathrm{GroupNorm}(\boldsymbol{F}_t^{l-1}))\big), \ \boldsymbol{F}_t^l = \boldsymbol{F}_t^{l-1} + \boldsymbol{R}_t^l, \ \tilde{\boldsymbol{F}}_t^l = \mathrm{Downsample}(\boldsymbol{F}_t^l) \tag{7}$$

Here, σ denotes the SiLU function, and $\mathbf{F}_t^l$ is passed to the next level after spatial downsampling. In the initial compression stage, this operation maps a feature tensor from $C \times H \times W$ to $4C \times (H/4) \times (W/4)$, reducing its feature volume to one quarter while increasing the channel capacity.

Unlike standard CNNs that treat inputs as static channels, our encoder integrates a ConvGRU-based gating mechanism within each hierarchical stage to model the temporal evolution of atmospheric states. Specifically, we employ a Convolutional GRU (ConvGRU) cell as the "convolutional operator".

Let $s \in \{1, 2\}$ index the states at $t-12$ and t, respectively, and let $\tilde{\mathbf{F}}_s^l$ denote the corresponding downsampled feature at scale l. We initialize $\mathbf{H}_0^l = 0$, and the ConvGRU updates its hidden state as:

$$\mathbf{r}_s^l = \sigma_g(\mathbf{W}_r^l * [\tilde{\mathbf{F}}_s^l, \mathbf{H}_{s-1}^l] + \mathbf{b}_r^l), \quad \mathbf{z}_s^l = \sigma_g(\mathbf{W}_z^l * [\tilde{\mathbf{F}}_s^l, \mathbf{H}_{s-1}^l] + \mathbf{b}_z^l), \tag{8}$$

$$\tilde{\mathbf{H}}_s^l = \tanh(\mathbf{W}_h^l * [\tilde{\mathbf{F}}_s^l, \mathbf{r}_s^l \odot \mathbf{H}_{s-1}^l] + \mathbf{b}_h^l), \quad \mathbf{H}_s^l = (1 - \mathbf{z}_s^l) \odot \mathbf{H}_{s-1}^l + \mathbf{z}_s^l \odot \tilde{\mathbf{H}}_s^l. \tag{9}$$

Here, $*$ denotes convolution, $\odot$ is the Hadamard product, and σ_g is the sigmoid function. After processing the two states, the final hidden state $\mathbf{H}_2^l$ encodes their temporal change at scale l. The resulting multi-scale features form the condition τ_θ^t:

$$\tau_\theta^t = \text{Embedding}(\mathbf{H}_2^1, \mathbf{H}_2^2, \ldots, \mathbf{H}_2^L) = \{\tau_{\theta,i}^t\}_{i=1}^L. \tag{10}$$

Each feature $\tau_{\theta,i}^t \in \mathbb{R}^{M_i \times d_c}$ is projected into the denoising block at the matching U-Net resolution. These features capture scale-dependent temporal variations more efficiently than the raw fields $\boldsymbol{X}_t$.

Denoiser with Attention mechanism Daily precipitation reflects accumulated effects of atmospheric processes acting across time and space (Trenberth et al., 2003). This nonlocal correspondence motivates spatial alignment between atmospheric conditions and daily precipitation features. We therefore introduce specialized conditioning to **maintain spatial consistency during denoising** and **bridge this cross-field gap**. We augment the U-Net backbone with a cross-attention mechanism (Vaswani et al., 2017; Jaegle et al., 2021a;b). At the i-th U-Net resolution, let $\boldsymbol{Z}_i = \varphi_i(\boldsymbol{Y}_t^k) \in \mathbb{R}^{N_i \times d_i}$ denote the flattened denoising feature and $\boldsymbol{C}_i = \tau_{\theta,i}^t \in \mathbb{R}^{M_i \times d_c}$ the corresponding atmospheric condition. Cross-attention is implemented as $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_a})V$, with

$$Q = \boldsymbol{Z}_i W_Q^{(i)}, \qquad K = \boldsymbol{C}_i W_K^{(i)}, \qquad V = \boldsymbol{C}_i W_V^{(i)}. \tag{11}$$

Here, d_a is the attention dimension, $W_Q^{(i)} \in \mathbb{R}^{d_i \times d_a}$, and $W_K^{(i)}, W_V^{(i)} \in \mathbb{R}^{d_c \times d_a}$ are learnable channel projections. This interaction allows precipitation features to selectively attend to atmospheric conditions at the matching spatial scale. Based on the above analysis, the Denoiser can predict the noise ϵ on the precipitation field. With Cross-attention, the prediction function can be expressed as:

$$\hat{\epsilon}_\theta = \text{Denoiser}(\boldsymbol{Y}_t^k, \tau_\theta^t, k) = f_{out}\left(\text{Cross-Attn}(\boldsymbol{Y}_t^k, \tau_\theta^t)\right) \tag{12}$$

The output function f_{out} restores the feature maps to the original resolution via the Sample Block and utilizes a convolutional layer to produce the predicted noise $\hat{\epsilon}_\theta$ with dimensions identical to $\boldsymbol{Y}_t^k$.

5 Experiments

5.1 Experimental Setup

Datasets We construct a domain-specific dataset integrating atmospheric analyses, observation-based precipitation, heterogeneous forecasts, and station measurements. Hourly **ERA5** states (Hersbach et al., 2020), using the variables and levels in Table 3, are paired by valid time with IMERG Final Run V07B 24-hour precipitation over $t-24, t$ (Pradhan et al., 2022). HRES and ENS are sourced from WeatherBench2 (Rasp et al., 2024), while GraphCast and Pangu-Weather forecasts are generated with their released models. All products are aligned to $0.25°$ and $1.5°$ grids. Data span 1998–2023, with 1998–2021 for training, 2023 for validation, and 2022 for testing. The 2022 test year matches WeatherBench2 forecast availability. Finally, 2,288 China Meteorological Administration stations are reserved for external evaluation of localized heavy-rainfall events.

Metrics Following (Bi et al., 2023; Gao et al., 2023; Sønderby et al., 2020), we quantify precipitation accuracy using the Anomaly Correlation Coefficient (**ACC**) and Root Mean Square Error (**RMSE**), and event skill using the Critical Success Index (**CSI**) and Heidke Skill Score (**HSS**). Probabilistic skill is evaluated with the Continuous Ranked Probability Score (**CRPS**) and Brier Score (**BS**). Deterministic results use one generated sample per case without sample selection. Metric definitions are provided in Section A.2 of the Appendix.

Baselines We evaluate our framework with the data-driven **Pangu-Weather** (Bi et al., 2023) and **GraphCast** (Lam et al., 2023) backbones, as well as the numerical **HRES** system (Haiden et al., 2021). Pangu-Weather provides no native precipitation output, while GraphCast and HRES enable direct comparison with their own precipitation forecasts. For probabilistic evaluation, we additionally include **GenCast** (Price et al., 2025) and **ENS** (Leutbecher et al., 2017) as external system-level baselines, since their ensembles also capture uncertainty in atmospheric trajectories.

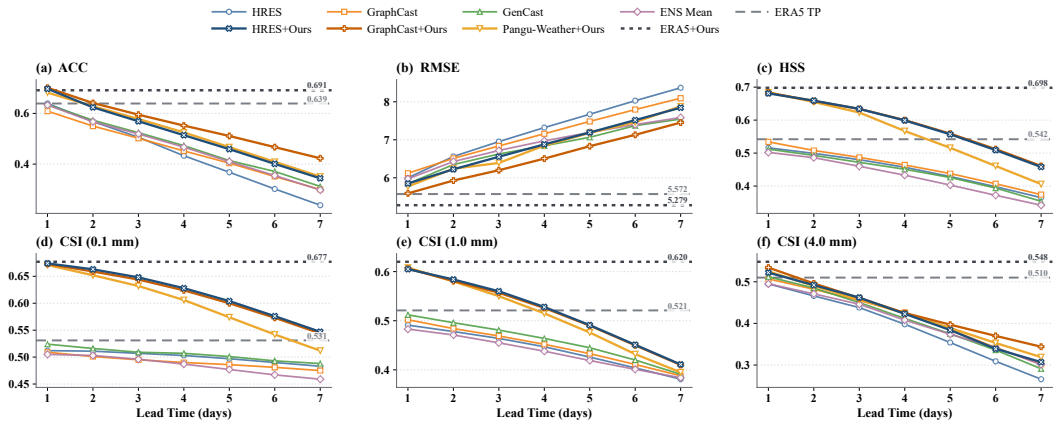


Figure 3: Global 1.5° deterministic skill over the 2022 test set for forecast backbones, their variants with our framework, and external system references. IMERG is used as the verification reference.

Implementation Our framework is implemented in PyTorch (Paszke et al., 2019) and trained for 300K steps. We use the AdamW optimizer (Kingma, 2014; Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-5} and $(\beta_1, \beta_2) = (0.9, 0.99)$. Following standard diffusion settings (Rombach et al., 2022), we use 1000 diffusion steps. During inference, DDIM (Song et al., 2021) reduces the denoising process to 100 steps. All experiments are conducted on four NVIDIA A800 GPUs. At both 0.25° and 1.5° resolutions, the resolution-specific checkpoint is applied to forecast states from all aligned backbones without backbone-specific retraining. For probabilistic evaluation, we generate $M = 50$ members from independently sampled Gaussian noise. The forecast-field fine-tuning diagnostic is described in Appendix A.3.

Table 1: Probabilistic forecasting for a 50-member ensemble.

Metrics	Models	Lead Time (Days) 1	2	3	4	5	6	7
CRPS ↓	ENS	1.528	1.740	1.824	1.975	2.161	2.346	2.484
	GenCast	1.512	1.651	1.778	1.943	2.147	2.255	2.368
	Pangu+**Ours**	1.483	1.619	1.756	1.910	2.061	2.208	2.336
	GraphCast+**Ours**	**1.477**	**1.569**	**1.660**	**1.769**	**1.898**	**2.035**	**2.180**
	HRES+**Ours**	1.525	1.637	1.747	1.867	2.006	2.156	2.302
BS ↓	ENS	0.193	0.201	0.211	0.229	0.237	0.239	0.245
	GenCast	0.177	0.195	0.209	0.216	0.230	0.240	0.247
	Pangu+**Ours**	0.141	0.149	0.164	0.180	0.200	0.221	0.240
	GraphCast+**Ours**	**0.138**	0.146	0.155	**0.164**	**0.181**	**0.199**	**0.218**
	HRES+**Ours**	0.140	**0.145**	**0.154**	0.166	0.185	0.203	0.219

5.2 Global Assessment at 1.5°

Deterministic Forecast We evaluate Pangu-Weather, GraphCast, and HRES equipped with our framework for lead times up to 7 days in Figure 3. Following Andrychowicz et al. (2023), CSI thresholds are set to $\{0.1, 1, 4\}$ mm. Based on these results, we have the following findings:

- Our framework enables Pangu-Weather to forecast precipitation and raises GraphCast's ACC by 0.103 while reducing its RMSE by 0.623 on average. GraphCast+Ours also outperforms GenCast. HRES+Ours performs comparably to GraphCast+Ours and better than Pangu-Weather+Ours, confirming that precipitation generation still depends on backbone atmospheric skill.
- ERA5+Ours serves as an analysis-conditioned diagnostic: its improvement over ERA5 total precipitation reflects closer agreement with IMERG, rather than an operational forecasting gain.
- HSS and CSI show consistent gains in precipitation-event discrimination across the thresholds. Improvements at 0.1 and 1.0 mm indicate better detection of small daily accumulations, while the 4-mm threshold supports improved detection of larger amounts within the evaluated range.

Probabilistic Forecast To investigate ensemble precipitation skill, we compare 50-member forecasts from our framework with **ENS** and **GenCast** in Table 1. From the results, we can find that:

- Overall, our variants are competitive with the external ensembles and achieve lower CRPS in most comparisons. GraphCast+Ours performs best at every reported lead time. At Day 7, its CRPS is 2.180, compared with 2.484 for ENS and 2.368 for GenCast. The lower CRPS indicates closer agreement between the predicted precipitation distributions and IMERG.
- All our variants achieve lower BS than ENS and GenCast at every lead time, indicating more accurate precipitation-event probabilities. At Day 7, GraphCast+Ours achieves 0.218, compared with 0.245 for ENS and 0.247 for GenCast, demonstrating stronger probabilistic event forecasting.

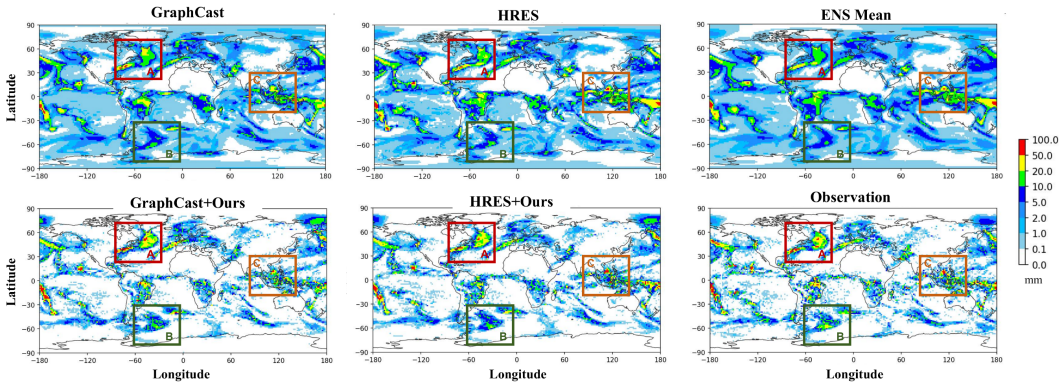


Figure 4: Spatial comparison of Day-3 precipitation forecasts from the baseline models, their variants with our framework, and the IMERG reference. Boxes A–C highlight regional differences.

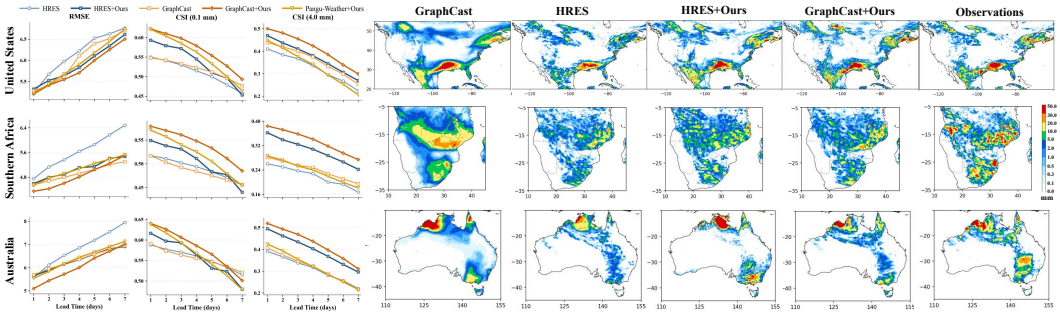


Figure 5: Regional evaluation over the United States, Southern Africa, and Australia. Left: RMSE and CSI across lead times. Right: precipitation fields from WFMs, their variants, and Observations.

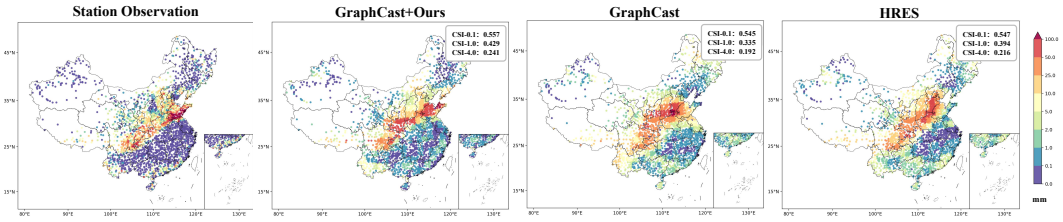


Figure 6: Station-based evaluation over North China in July 2022, initialized on June 29.

Visualization We visualize Day-3 precipitation in Figure 4. All systems capture the large-scale patterns, while the base forecasts are smoother at local scales. As expected, averaging ENS members further suppresses member-level variability. With our framework, forecasts show more distinct local accumulation maxima. In Boxes A and C, their spatial patterns and daily amounts better match IMERG, while Box B shows clearer rainband boundaries. These examples support improved spatial representation of daily precipitation, although visual sharpness alone does not establish forecast accuracy. More visualizations are shown in Figure 11 of the Appendix.

5.3 Regional Evaluation at 0.25°

Region Comparisons. Figure 5(left) evaluates our framework over the United States, Southern Africa, and Australia at 0.25°. Across these regions, both HRES+Ours and GraphCast+Ours improve precipitation forecasting, with GraphCast+Ours achieving the best overall performance. These results demonstrate robust coupling with aligned 0.25° WFM and NWP outputs and effective use of their fine-scale atmospheric structures for regional precipitation prediction. Figure 5(right) compares 24-h precipitation forecasts. GraphCast oversmooths rainfall, while HRES misses local maxima despite capturing large-scale patterns. Our framework preserves this organization and recovers sharper, heterogeneous daily rainfall patterns over Southern Africa and northern Australia. Closer agreement with IMERG further indicates that our framework better represents localized daily accumulation patterns at 0.25°.

Case Study at 2,288 Stations We examine North China's early-July 2022 heavy-rainfall event, where Chaba's remnants and a westerly trough organized moisture transport and a broad rainband

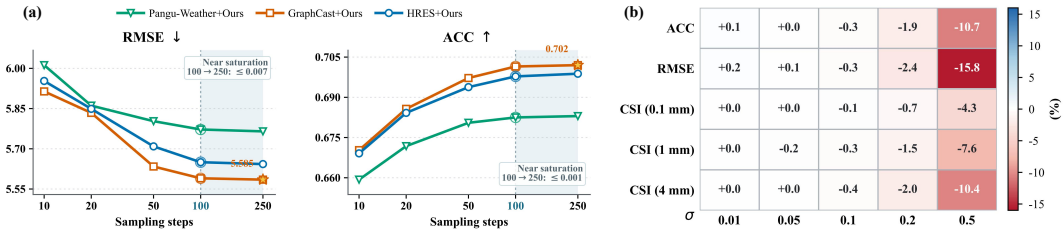


Figure 7: (a) Impact of sampling steps. RMSE and ACC approach saturation at 100 steps. (b) Relative performance changes under Gaussian perturbed inputs; red indicates degradation.

with local maxima. In Figure 6, HRES and GraphCast capture the rainband but smooth contrasts in daily totals. GraphCast+Ours preserves its placement and recovers accumulation maxima without domain-wide intensification. Multi-threshold CSI gains support improved event detection.

5.4 Mechanism and Robustness Analysis

Effectiveness of the Model Design Table 2 compares partial and complete configurations under the same data split and evaluation protocol. **(i) Temporal extraction.** Among the temporal configurations, ConvGRU consistently outperforms ConvLSTM under both conditioning backbones, supporting its use for compact atmospheric evolution encoding. **(ii) Spatial alignment.** Among the alignment configurations, cross-attention performs better than gated fusion, showing the benefit of selectively aligning atmospheric and precipitation features. **(iii) Complete-model comparison.** U-Net remains competitive for low-intensity event detection but is less accurate for the overall intensity distribution, while CorrDiff's lower scores are consistent with a mismatch between its coarse-to-fine objective and this same-resolution generation setting. The full configuration performs best across the reported metrics, indicating complementary gains from temporal extraction and spatial alignment.

Table 2: Day-1 configuration comparison. "–" denotes an omitted proposed module; U-Net and CorrDiff replace the complete generator.

Condition	Temp	Align	Generator	ACC↑	RMSE↓	HSS↑	$CSI_{0.1}$ ↑	CSI_1 ↑	CSI_4 ↑
GraphCast	–	–	Native	0.6091	6.120	0.5347	0.5097	0.4928	0.4271
	GRU	–	Diff.	0.6217	5.924	0.5914	0.5612	0.5267	0.4826
	LSTM	–	Diff.	0.5942	6.318	0.5165	0.4942	0.4725	0.3945
	–	Attn.	Diff.	0.6571	5.721	0.6213	0.6158	0.5572	0.5185
	–	Gated	Diff.	0.6126	6.067	0.5678	0.5367	0.5126	0.4438
	–	–	U-Net	0.4071	9.840	0.6268	0.6458	0.5464	0.4535
	–	–	CorrDiff	0.5004	6.939	0.5733	0.5877	0.5067	0.4263
	GRU	Attn.	Diff. (Ours)	**0.7015**	**5.590**	**0.6832**	**0.6736**	**0.6061**	**0.5340**
HRES	–	–	Native	0.6394	6.010	0.5164	0.5126	0.4913	0.4151
	GRU	–	Diff.	0.6385	5.984	0.5927	0.5581	0.5357	0.4782
	LSTM	–	Diff.	0.5905	6.359	0.5118	0.4909	0.4769	0.3916
	–	Attn.	Diff.	0.6572	5.841	0.6174	0.6087	0.5581	0.5108
	–	Gated	Diff.	0.6083	6.132	0.5541	0.5318	0.5118	0.4397
	–	–	U-Net	0.4023	9.733	0.6267	0.6451	0.5479	0.4521
	–	–	CorrDiff	0.4911	7.095	0.5719	0.5888	0.5064	0.4233
	GRU	Attn.	Diff. (Ours)	**0.6978**	**5.650**	**0.6818**	**0.6742**	**0.6054**	**0.5221**

Efficiency and Robustness Analysis As shown in Figure 7(a), increasing the sampling steps from 10 to 100 consistently improves RMSE and ACC across all three backbones. Beyond 100 steps, the performance curves largely saturate, with only marginal gains up to 250 steps. We therefore use 100 sampling steps by default, providing a favorable trade-off between forecast quality and computational cost. In addition, we perturb normalized atmospheric inputs with Gaussian noise of standard deviation σ. As shown in Figure 7(b), performance changes remain limited for small perturbations and degrade gradually as σ increases, indicating controlled sensitivity to synthetic input errors.

6 Conclusion

We present *Dynamics–Precipitation Decoupling* for daily precipitation generation at medium-range lead times. Trained on ERA5–IMERG pairs, the diffusion generator maps non-precipitation atmospheric states to daily accumulations through recurrent encoding and cross-field alignment. It does not use the backbone's precipitation outputs. One generator per resolution is reused across aligned backbones without backbone-specific retraining. Global and regional evaluations show improved deterministic skill, with competitive probabilistic performance in the global evaluation. A heavy-rainfall case study at 2,288 stations further supports improved daily accumulation patterns and event detection. Ablations support complementary contributions from temporal encoding and spatial alignment. The framework samples precipitation conditional on a fixed atmospheric trajectory, and its skill remains dependent on the backbone forecast. Further evaluation across forecasting systems and heavy-rainfall events remains important. These results demonstrate the potential of decoupled precipitation generation as a reusable component of medium-range forecasting systems.

REFERENCES

Daniel Abdi, Isidora Jankov, Paul Madden, Vanderlei Vargas, Timothy A Smith, Sergey Frolov, Montgomery Flora, and Corey Potvin. Hrrrcast: a data-driven emulator for regional weather forecasting at convection allowing scales. *arXiv preprint arXiv:2507.05658*, 2025.

Sojung An, Tae-Jin Oh, Eunha Sohn, and Donghyun Kim. Deep learning for precipitation nowcasting: A survey from the perspective of time series forecasting. *Expert Systems with Applications*, 268: 126301, 2025.

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Marcin Andrychowicz, Lasse Espeholt, Di Li, Samier Merchant, Alexander Merose, Fred Zyda, Shreya Agrawal, and Nal Kalchbrenner. Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*, 2023.

Shejule Priya Ashok and Sreeja Pekkat. A systematic quantitative review on the performance of some of the recent short-term rainfall forecasting techniques. *Journal of Water and Climate Change*, 13 (8):3004–3029, 2022.

Georgy Ayzel, Tobias Scheffer, and Maik Heistermann. Rainnet v1. 0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644, 2020.

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

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

Peter Bauer, Peter D Dueben, Torsten Hoefler, Tiago Quintino, Thomas C Schulthess, and Nils P Wedi. The digital revolution of earth-system science. *Nature Computational Science*, 1(2):104–113, 2021.

SARAH Benziane and USTO MB. Survey: Rainfall prediction precipitation, review of statistical methods. *WSEAS Transactions on Systems*, 23:47–59, 2024.

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.

Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu, Bo Lu, Deliang Chen, Shang-Ping Xie, Libo Wu, Qingchen Chao, Chensen Lin, et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1):6425, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, et al. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1):5145, 2022.

Zhihan Gao, Xingjian Shi, Hao Wang, Dit-Yan Yeung, Wang-chun Woo, and Wai-Kin Wong. Deep learning and the weather forecasting problem: Precipitation nowcasting. *Deep Learning for the Earth Sciences: a comprehensive approach to remote sensing, climate science, and geosciences*, pp. 218–239, 2021.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

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*, 36:78621–78656, 2023.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

T Haiden, M Janousek, F Vitart, Z Ben-Bouallegue, L Ferranti, F Prates, and D Richardson. Evaluation of ecmwf forecasts, including the 2021 upgrade, ecmwf technical memoranda 884. *Reading, UK: European Centre for Medium Range Weather Forecasts*, 2021.

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.

Timothy David Hewson and Fatima Maria Pillosu. A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, 2(1):132, 2021.

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

Arthur Y Hou, Ramesh K Kakar, Steven Neeck, Ardeshir A Azarbarzin, Christian D Kummerow, Masahiro Kojima, Riko Oki, Kenji Nakamura, and Toshio Iguchi. The global precipitation measurement mission. *Bulletin of the American meteorological Society*, 95(5):701–722, 2014.

Abba Ibrahim, Aimrun Wayayok, Helmi Zuhaidi Mohd Shafri, Noorellimia Mat Toridi, and Idris Wada Muhammad. A two-stage bias-correction and super-resolution framework for post-processing climate model outputs. *EGUsphere*, 2026:1–26, 2026.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021a.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021b.

Donglai Jiao, Nannan Xu, Fan Yang, and Ke Xu. Evaluation of spatial-temporal variation performance of era5 precipitation data in china. *Scientific Reports*, 11(1):17956, 2021.

DP Kingma. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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*, 632(8027):1060–1066, 2024.

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.

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.

Martin Leutbecher, Sarah-Jane Lock, Pirkka Ollinaho, Simon TK Lang, Gianpaolo Balsamo, Peter Bechtold, Massimo Bonavita, Hannah M Christensen, Michail Diamantakis, Emanuel Dutra, et al. Stochastic representations of model uncertainties at ecmwf: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143(707):2315–2339, 2017.

Qi Liu, Xiao Lou, Zhongwei Yan, Yajie Qi, Yuchao Jin, Shuang Yu, Xiaoliang Yang, Deming Zhao, and Jiangjiang Xia. Deep-learning post-processing of short-term station precipitation based on nwp forecasts. *Atmospheric Research*, 295:107032, 2023.

Yuhao Liu, James Doss-Gollin, Qiushi Dai, Ashok Veeraraghavan, and Guha Balakrishnan. Downscaling extreme precipitation with wasserstein regularized diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, pp. 25904–25938. PMLR, 2023.

Tim Palmer. The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, 9(2):730–734, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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.

Jaideep Pathak, Yair Cohen, Piyush Garg, Peter Harrington, Noah Brenowitz, Dale Durran, Morteza Mardani, Arash Vahdat, Shaoming Xu, Karthik Kashinath, et al. Kilometer-scale convection-allowing model emulation using generative diffusion modeling. *Science Advances*, 12(5):eadv0423, 2026.

Rajani K Pradhan, Yannis Markonis, Mijael Rodrigo Vargas Godoy, Anahí Villalba-Pradas, Konstantinos M Andreadis, Efthymios I Nikolopoulos, Simon Michael Papalexiou, Akif Rahim, Francisco J Tapiador, and Martin Hanel. Review of gpm imerg performance: A global perspective. *Remote Sensing of Environment*, 268:112754, 2022.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NIPS*, 28, 2015.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*, 2020.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Qiaohong Sun, Chiyuan Miao, Qingyun Duan, Hamed Ashouri, Soroosh Sorooshian, and Kuo-Lin Hsu. A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of geophysics*, 56(1):79–107, 2018.

Ye Tian, Yan Ji, Xichao Gao, Xing Yuan, and Xiefei Zhi. Post-processing of short-term quantitative precipitation forecast with the multi-stream convolutional neural network. *Atmospheric Research*, 309:107584, 2024.

Kevin E Trenberth, Aiguo Dai, Roy M Rasmussen, and David B Parsons. The changing character of precipitation. *Bulletin of the American Meteorological Society*, 84(9):1205–1218, 2003.

Hristos Tyralis and Georgia Papacharalampous. A review of predictive uncertainty estimation with machine learning. *Artificial Intelligence Review*, 57(4):94, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elisabeth Vogel, Markus G Donat, Lisa V Alexander, Malte Meinshausen, Deepak K Ray, David Karoly, Nicolai Meinshausen, and Katja Frieler. The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, 14(5):054010, 2019.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10): e2021MS002954, 2022.

Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

Wenhao Xie, Shanzhen Yi, Chuang Leng, Defeng Xia, Mingli Li, Zewen Zhong, and Jianfeng Ye. The evaluation of imerg and era5-land daily precipitation over china with considering the influence of gauge data bias. *Scientific Reports*, 12(1):8085, 2022.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Janni Yuval, Ian Langmore, Dmitrii Kochkov, and Stephan Hoyer. Neural general circulation models for modeling precipitation. *Science Advances*, 12(2):eadv6891, 2026.

Tuantuan Zhang, Zhongmin Liang, Chenglin Bi, Jun Wang, Yiming Hu, and Binquan Li. Statistical post-processing for precipitation forecast through deep learning coupling large-scale and local-scale spatiotemporal information. *Water Resources Management*, 39(1):145–160, 2025.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619 (7970):526–532, 2023.

Honglin Zhu, Huizeng Liu, Qiming Zhou, and Aihong Cui. Towards an accurate and reliable downscaling scheme for high-spatial-resolution precipitation data. *Remote Sensing*, 15(10):2640, 2023.

Honglin Zhu, Qiming Zhou, and Jukka M Krisp. Exploring machine learning approaches for precipitation downscaling. *Geo-Spatial Information Science*, pp. 1–17, 2025.

ETHICS STATEMENT AND LIMITATIONS

Cross-backbone reuse is empirically validated only for the aligned forecast systems considered in this work, while the $1.5°$ and $0.25°$ settings require separately trained models. Training on ERA5 states also introduces a distribution shift when conditioning on imperfect forecasts, which we probe through noise perturbation and forecast-field fine-tuning experiments. Future work should further evaluate transfer across other reanalyses, such as JRA and MERRA-2, and operational products such as GFS. IMERG retrieval uncertainty and the additional cost of diffusion sampling remain further limitations. All datasets are publicly available and contain no personal information; nevertheless, generated precipitation should be used as decision support and requires regional validation, expert oversight, and clear communication of forecast uncertainty.

AI USE STATEMENT

In this work, we used LLMs as an auxiliary tool to polish the writing and improve the manuscript's clarity. Specifically, we use it to review the entire text for grammatical accuracy, improve sentence structure, and ensure consistent phrasing and tone throughout the paper.

A APPENDIX

A.1 BASELINES

- **Pangu-Weather** Developed by Huawei Cloud, Pangu-Weather is the first AI model to surpass traditional NWP accuracy. It uses four 3D Earth-specific Transformer models for 1-, 3-, 6-, and 24-hour forecasts, each with approximately 64 million parameters (256 million in total), and combines them through hierarchical temporal aggregation. The model maintains global physical consistency while delivering massive speedups in inference compared to traditional numerical systems.
- **GraphCast** Developed by Google DeepMind, GraphCast is a mid-range forecasting model based on Graph Neural Networks (GNNs). It employs a "Grid-to-Mesh-to-Grid" architecture, mapping atmospheric variables onto a multi-scale icosahedral mesh to capture long-range dependencies and resolve polar singularities. At 0.25° resolution, it delivers SOTA deterministic accuracy across thousands of meteorological variables.
- **GenCast** As an advanced global deep learning forecasting model, GenCast can generate probabilistic forecasts that surpass those of traditional ensemble systems. It demonstrates superior performance in providing high-resolution, physically consistent precipitation distribution and extreme weather warnings by efficiently capturing the complex uncertainties of atmospheric evolution.
- **HRES** Operated by ECMWF, HRES is the world's leading Numerical Weather Prediction (NWP) system and the industry's "gold standard". It describes atmospheric evolution through complex fluid dynamics and physical parameterizations, integrated with massive global observations. Despite the rise of AI models, HRES remains the essential benchmark for validating physical reliability in weather forecasting.
- **ENS** The Ensemble Prediction System (ENS) operated by ECMWF is designed to quantify atmospheric uncertainty by generating multiple forecast members through subtle perturbations in initial conditions or model parameters. Unlike the single deterministic HRES, ENS enhances medium-range robustness via the ensemble mean (ENS-Mean) or probabilistic distributions. In precipitation forecasting, it serves as a critical operational benchmark for evaluating AI models' proficiency in uncertainty quantification and probabilistic accuracy.

A.2 METRICS

Root mean square error (RMSE) Root Mean Square Error (RMSE) evaluates the average size of the errors in a dataset. It measures how well the predicted values align with the true values, where a

Table 3: ERA5 variable–level interface used in the experiments.

Type	Variable	Abbreviation	Levels
Static	Land–sea mask	LSM	–
Static	Orography	ORG	–
Surface	2-m temperature	T2m	–
Surface	10-m zonal wind	U10	–
Surface	10-m meridional wind	V10	–
Surface	Mean sea-level pressure	MSLP	–
Upper	Geopotential	Z	50, 100, 150
Upper	U component of wind	U	200, 250, 300
Upper	V component of wind	V	400, 500, 600
Upper	Temperature	T	700, 850
Upper	Specific humidity	Q	925, 1000

smaller RMSE indicates better prediction accuracy.

$$\mathrm{RMSE} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} L(h)(\hat{\boldsymbol{Y}}_{t,h,w} - \boldsymbol{Y}_{t,h,w})^2} \tag{13}$$

where $L(h)$ is introduced as a latitude-based weighting factor:

$$L(h) = \frac{\cos(\mathrm{lat}(h))}{\frac{1}{H}\sum_{h=1}^{H}\cos(\mathrm{lat}(h))} \tag{14}$$

Anomaly correlation coefficient (ACC) The Anomaly Correlation Coefficient (ACC) assesses forecast accuracy by quantifying the correlation between predicted and observed anomalies. It is widely used in climate and weather prediction to assess how effectively the model identifies deviations from the climatological mean. A higher ACC value signifies better forecasting performance.

$$\mathrm{ACC} = \frac{\sum_{t,h,w} L(h)(\hat{\boldsymbol{Y}}_{t,h,w} - C_{t,h,w})(\boldsymbol{Y}_{t,h,w} - C_{t,h,w})}{\sqrt{\sum_{t,h,w} L(h)(\hat{\boldsymbol{Y}}_{t,h,w} - C_{t,h,w})^2 \sum_{t,h,w} L(h)(\boldsymbol{Y}_{t,h,w} - C_{t,h,w})^2}} \tag{15}$$

where climatology C is computed only from the 1998–2021 training period. The calculation of $L(h)$ follows Eq. 14.

Critical Success Index (CSI) CSI measures the overlap between observed and predicted precipitation events and ranges from 0 (worst) to 1 (best). For global evaluation, contingency counts are pooled over test cases and grid cells using latitude weights.

$$\mathrm{CSI} = \frac{TP}{TP + FP + FN} \tag{16}$$

where the weighted contingency counts are defined as:

$$TP = \sum_{t,h,w} L(h)\,\mathbb{I}[(\hat{\boldsymbol{Y}}_{t,h,w} \geq th) \wedge (\boldsymbol{Y}_{t,h,w} \geq th)], \tag{17}$$

$$FP = \sum_{t,h,w} L(h)\,\mathbb{I}[(\hat{\boldsymbol{Y}}_{t,h,w} \geq th) \wedge (\boldsymbol{Y}_{t,h,w} < th)], \tag{18}$$

$$FN = \sum_{t,h,w} L(h)\,\mathbb{I}[(\hat{\boldsymbol{Y}}_{t,h,w} < th) \wedge (\boldsymbol{Y}_{t,h,w} \geq th)], \tag{19}$$

$$TN = \sum_{t,h,w} L(h)\,\mathbb{I}[(\hat{\boldsymbol{Y}}_{t,h,w} < th) \wedge (\boldsymbol{Y}_{t,h,w} < th)] \tag{20}$$

Here, $\mathbb{I}(\cdot)$ is an indicator and $th \in \{0.1, 1.0, 4.0\}$ mm is the precipitation threshold.

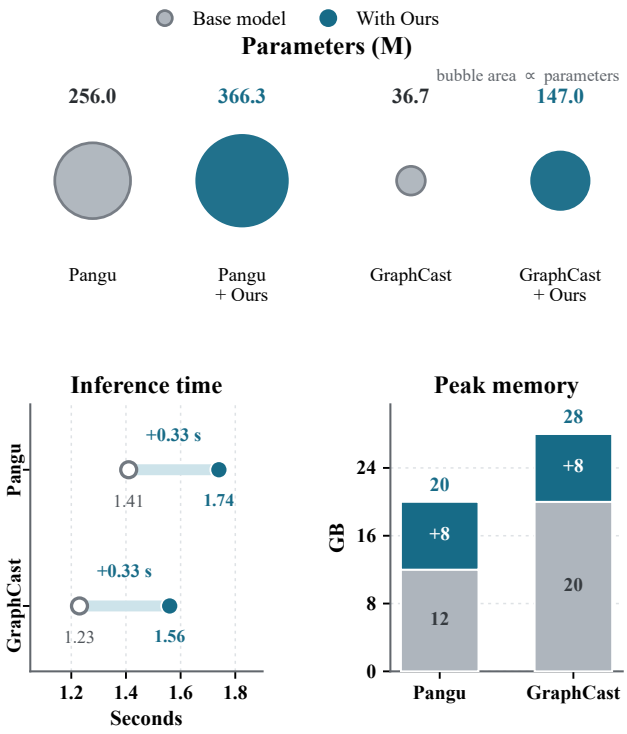


Figure 8: Computational overhead of our framework on Pangu-Weather and GraphCast in terms of parameters, inference time, and peak GPU memory. Pangu-Weather parameter counts include all four forecast models.

Heidke Skill Score (HSS) HSS measures the accuracy of a binary forecast relative to random chance. It accounts for true positives, true negatives, false positives, and false negatives; higher values indicate better event discrimination.

$$\mathrm{HSS} = \frac{2\,(TP \times TN - FP \times FN)}{(TP + FN)\,(FN + TN) + (TP + FP)\,(FP + TN)} \tag{21}$$

In this paper, we use a precipitation threshold $th = 0.1\,\mathrm{mm}$ to classify rain versus no-rain events.

Continuous Ranked Probability Score (CRPS) is estimated for our ensembles as follows:

$$\mathrm{CRPS}_{t,h,w} = \frac{1}{M}\sum_{m=1}^{M} |\hat{y}_{m,t,h,w} - y_{t,h,w}| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} |\hat{y}_{m,t,h,w} - \hat{y}_{n,t,h,w}| \tag{22}$$

where $M = 50$ for all compared systems and $\hat{y}_{m,t,h,w}$ is the m-th ensemble member at a given case and grid point. Following Price et al. (2025), we use the traditional estimator under the matched ensemble size. Gridpoint CRPS values are averaged over the evaluation cases and spatial domain using the latitude weights in Eq. 14.

Brier Score (BS) BS measures the mean squared difference between predicted probabilities and binary outcomes for precipitation events. Given an ensemble of M members $\hat{Y} = \{\hat{y}^{(1)}, \hat{y}^{(2)}, \ldots, \hat{y}^{(M)}\}$ and a specific rainfall threshold th, we first define the empirical predictive probability P at each grid point and time step as the fraction of members exceeding the threshold:

$$P = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}(\hat{y}^{(m)} \geq th) \tag{23}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The Brier Score is then calculated by averaging the squared error between this probability and the binary observation $O = \mathbb{I}(y \geq th)$:

$$\mathrm{BS} = \frac{1}{T \cdot H \cdot W}\sum_{t,h,w} L(h)(P_{t,h,w} - O_{t,h,w})^2 \tag{24}$$

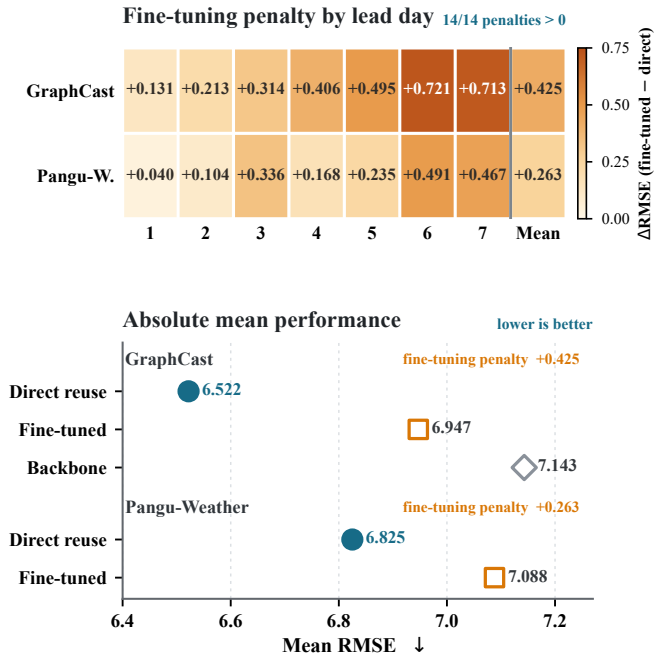


Figure 9: Cross-backbone reuse and fine-tuning diagnostic.

where T, H, and W denote the number of evaluated cases, grid height, and grid width. BS ranges from 0 to 1, with 0 indicating perfect event-probability accuracy; we use it as an aggregate skill score rather than attributing changes separately to calibration or resolution. For BS, the precipitation threshold is set to $0.1\mathrm{mm}$.

A.3 Additional Results

Complexity and Efficiency Analysis Figure 8 reports the computational overhead of the proposed generator. It introduces 110.3M parameters, 0.33 s per generated sample, and approximately 8 GB of additional peak GPU memory under the reported configuration, making the relative overhead more pronounced for lightweight backbones. This non-negligible cost mainly arises from iterative diffusion sampling. However, the generator is trained once and reused across different forecast backbones at the same resolution, avoiding backbone-specific retraining, and fewer sampling steps can be used when lower latency is preferred. Together with the sampling-step analysis, these results expose the trade-off between precipitation quality and computational cost.

Temporal-Density Ablation. We compare the default two-state input $(\hat{\boldsymbol{X}}_{t-12}, \hat{\boldsymbol{X}}_t)$ with four states sampled at 6-hour intervals. The two configurations achieve mean RMSEs of 6.520 and 6.518, respectively, compared with 7.143 for native GraphCast. Denser conditioning reduces RMSE by only 0.002 (0.03%), while the two-state input reduces inference overhead from 0.84 s to 0.33 s under the evaluated configuration. We therefore adopt two-state conditioning for its lower computational cost and similar forecast accuracy.

Cross-Backbone Reuse and Fine-Tuning Diagnostic At a fixed resolution, GraphCast and Pangu-Weather are coupled with the same ERA5-trained generator checkpoint and use an identical interface in terms of variables, grids, units, and normalization. This controlled setting isolates the effect of the generator adaptation strategy from differences in model configuration. Direct reuse applies the pretrained generator to forecast states without parameter updates, whereas the diagnostic fine-tuning variant starts from the same checkpoint and updates only the generator using forecast-state–IMERG pairs; the forecast backbone remains frozen.

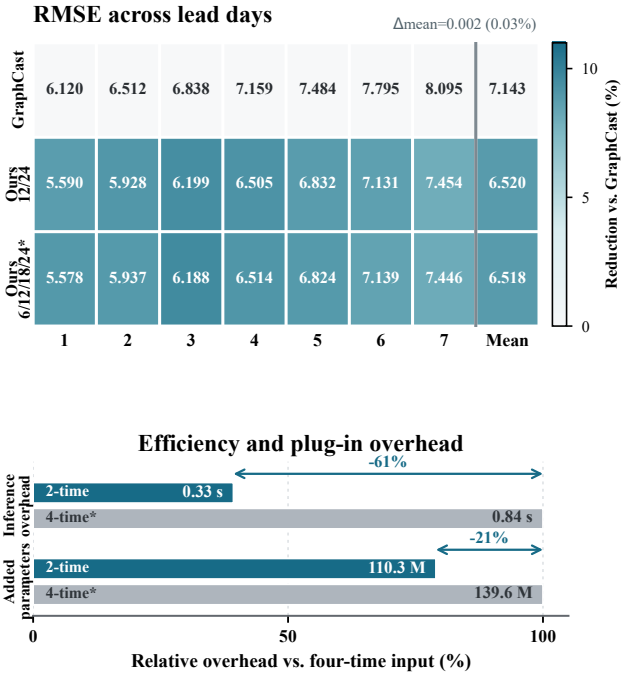


Figure 10: Temporal-density ablation under GraphCast conditioning: forecast accuracy and computational overhead with two-state and four-state atmospheric inputs.

As shown in Figure 9, fine-tuning increases RMSE at every evaluated lead day for both backbones. The mean RMSE increases from 6.522 to 6.947 for GraphCast, corresponding to a fine-tuning penalty of 0.425, and from 6.825 to 7.088 for Pangu-Weather, corresponding to a penalty of 0.263. The degradation becomes particularly pronounced at longer lead times, where forecast errors and their lead-time dependence are stronger. Moreover, direct reuse with GraphCast achieves a substantially lower mean RMSE than the original GraphCast forecast (6.522 versus 7.143). These consistent results suggest that the ERA5-trained checkpoint captures a transferable atmosphere–precipitation relationship, while fitting the generator to backbone-specific forecast errors may cause it to absorb lead-time-dependent biases and consequently weaken this learned relationship. We therefore adopt direct cross-backbone reuse in the primary method and include fine-tuning only as a diagnostic experiment.

Additional global visualization comparison. We provide additional visualization comparisons for all models at lead times of 1, 3, 5, and 7 days in Figure 11.

Additional Station Observation Comparison. We have added more site observation visualizations and comparisons at different time points in Figure 12.

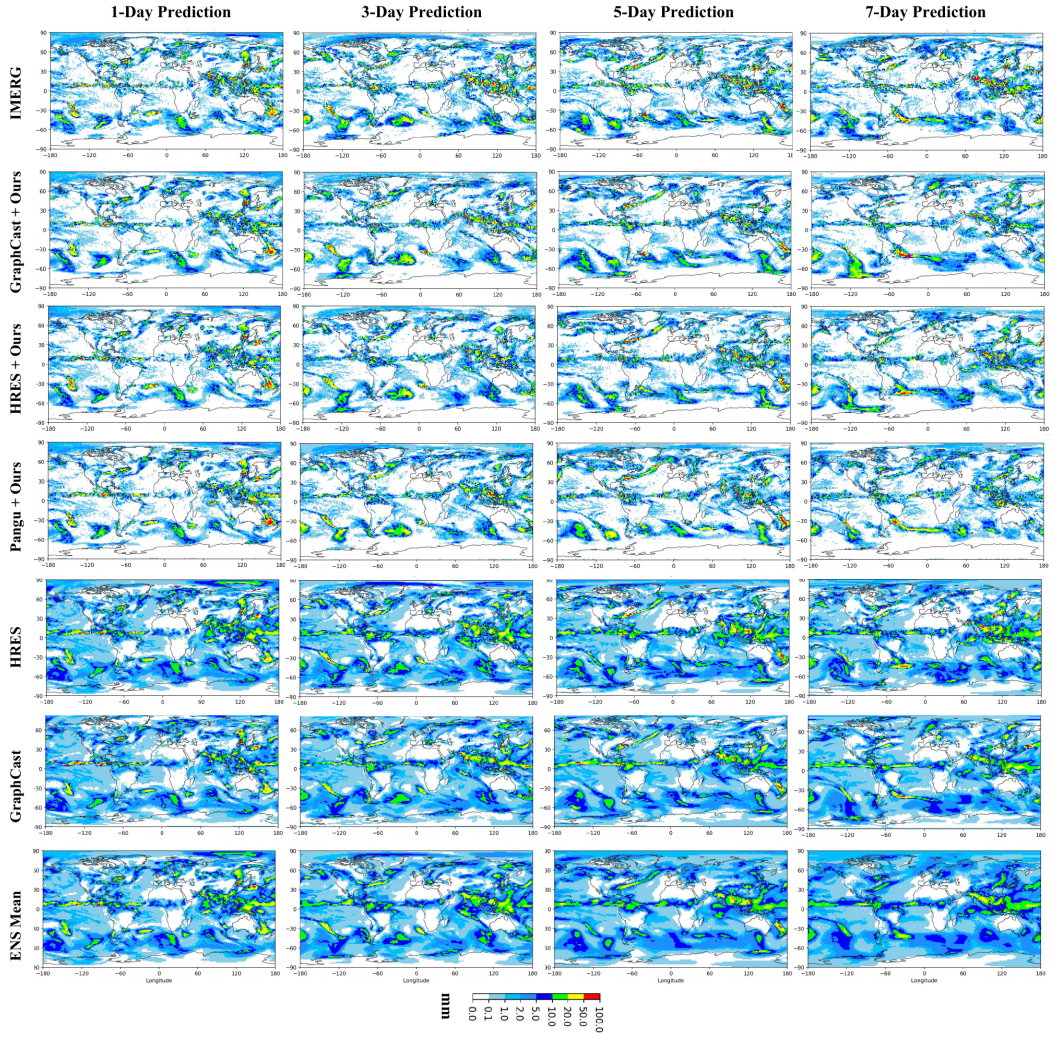


Figure 11: Global precipitation fields at lead times of 1, 3, 5, and 7 days, with IMERG as the reference. Forecasts are initialized at 00:00 UTC on July 6, 2022.

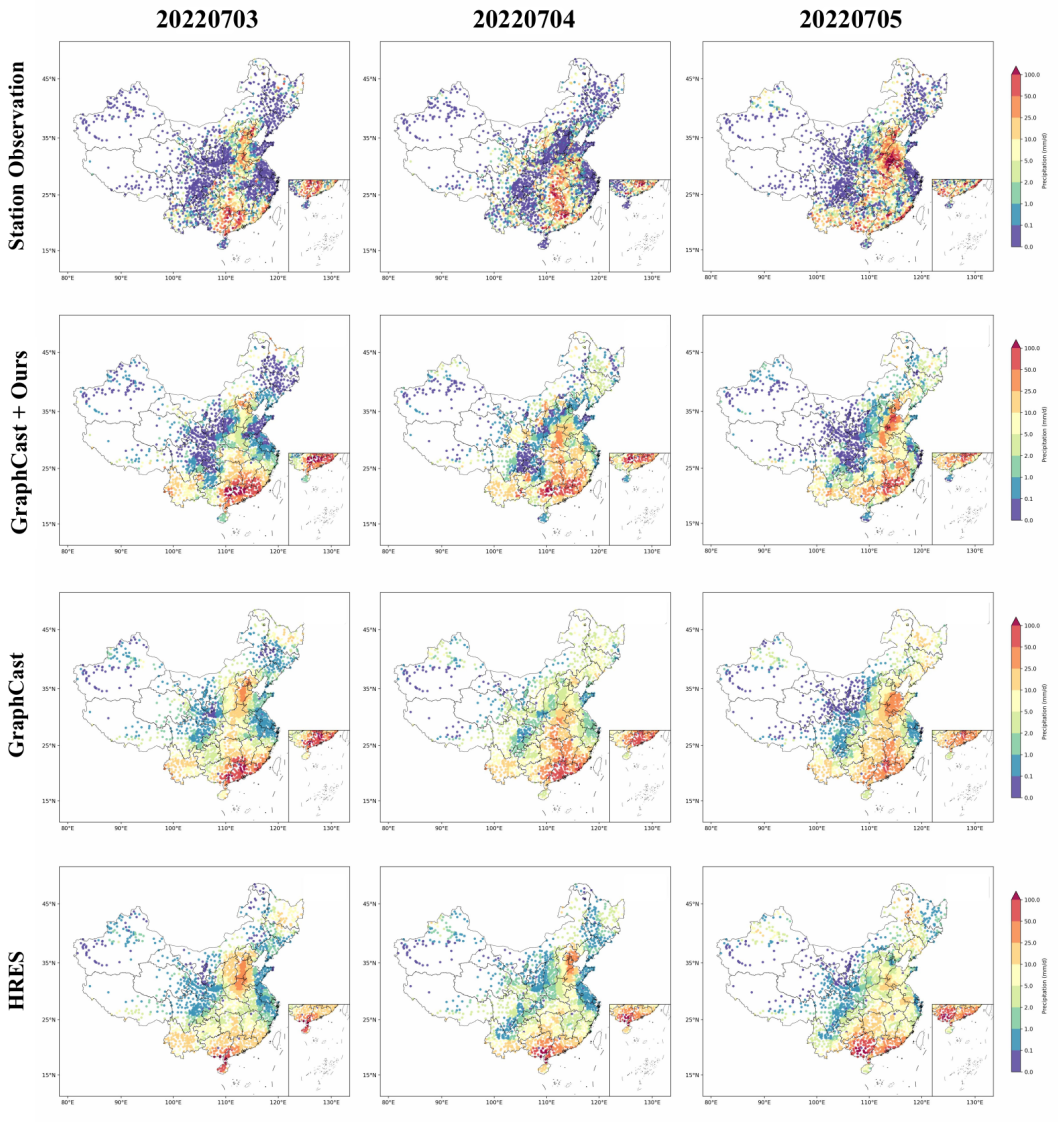


Figure 12: Station-based comparison across three valid times, with forecasts initialized on June 29, 2022.