DIAGNOSING PRECIPITATION WITH AN GRAPH LIKE ATTENTION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate diagnosis of fine-scale precipitation patterns remains a critical yet underexplored task in Earth system modeling. Unlike sequence-to-sequence forecasting, spatial diagnostics demand high spatial fidelity, robustness to intermittency, and interpretability—especially under extreme weather conditions. In this study, we propose a deep learning framework that operates entirely in the spatial domain, without relying on temporal encoding. The model begins with a pseudo spectral-space transformation to reorganize spatial modes, followed by cluster-wise attention that enables localized, structure-aware feature fusion. A sparse channel selection module further improves compactness and generalization. We train and evaluate our method on ensemble forecast fields and corresponding daily precipitation observations from 2010 to 2018, focusing on daily accumulated precipitation. Compared to strong baselines including LetNet, UNet with SSIM loss, FourCastNet, and SwinUnet, our method achieves consistently superior performance across standard metrics such as RMSE, R^2, Pearson correlation, Hanssen–Kuipers (HK) score, and multi-threshold Equitable Threat Score (ETS). Notably, the model demonstrates improved skill in detecting extreme precipitation events. Spatial and per-sample analyses further confirm its accuracy and robustness. These results highlight the promise of structure-aware spatial models for operational precipitation diagnostics and post-processing tasks.

1 INTRODUCTION

Precipitation constitutes a fundamental element of the hydrological cycle and serves as a critical determinant of weather-related hazards. While numerical weather prediction (NWP) and machine learning–based forecasting systems have advanced significantly, precipitation forecasting continues to present substantial challenges owing to its pronounced spatial intermittency, steep gradients, and dependence on subgrid-scale physical processes Nogueira (2014); Laprise (2008). Conventional global weather models face inherent limitations in direct precipitation prediction due to resolution constraints and parameterization uncertainties. This has fostered the development of an alternative research direction—precipitation diagnosis—which aims to derive precipitation fields from more robust and predictable atmospheric variables, including temperature, humidity, pressure, and vertical velocity Bourgouin (2000). FourCastNet Pathak et al. (2022), for instance, specialized precipitation neural modules perform data-driven mapping between predicted physical fields and precipitation estimates. Such diagnostic modules generally function without temporal recurrence and can be trained independently of the forecasting model, facilitating modular implementation.

These spatial diagnostic approaches are distinct from short-term precipitation estimation techniques employing radar or satellite observations Zhang et al. (2023); Espeholt et al. (2022). Nowcasting systems predominantly utilize high-frequency observational data and concentrate on very short lead times (generally less than six hours), whereas our methodology processes coarser temporal resolution data and utilizes model outputs rather than direct measurements as inputs. This differentiation is crucial: our objective centers on establishing a physically constrained mapping between globally modeled atmospheric variables and precipitation, rather than pursuing short-term tracking of convective systems.

In this work, we address the problem of precipitation diagnosis by focusing on its inherently anisotropic, intermittent, and multi-scale spatial structure. Unlike traditional approaches that rely

on autoregressive rollout or assume stationary global patterns, we explore a representation-driven perspective. Specifically, we recast the diagnostic task as a structured spatial inference problem, leveraging graph-based token interaction, domain-informed spatial clustering, and adaptive sparsification to improve accuracy and generalization under diverse meteorological regimes.

To facilitate the development and evaluation of our diagnostic framework, we employ the 50-member ensemble forecasts Park et al. (2008) from the ECMWF Ensemble Prediction System (ENS) Roberts et al. (2018) as model inputs and use the Tropical Rainfall Measuring Mission (TRMM) Simpson et al. (1996) daily precipitation product as observational reference. ENS provides a probabilistic representation of forecast uncertainty by perturbing initial conditions, which is particularly important for precipitation due to its inherently stochastic and highly nonlinear nature. Meanwhile, TRMM offers high-quality, observation-based daily precipitation estimates over the tropics and subtropics, making it a suitable benchmark for diagnosing spatial precipitation patterns at synoptic scales. This dataset pairing allows us to investigate the diagnostic mapping under realistic forecast variability while leveraging reliable ground-truth observations.

In this study, we propose a novel spatial diagnostic model for precipitation fields, designed to address challenges posed by spatial intermittency, non-stationarity, and multi-scale variability. Instead of relying on spatiotemporal rollouts or assumptions of globally stationary patterns, our architecture employs a pseudo-spectral inspired spatial encoding, structured token interaction through graph-based spatial clustering and learned sparsification. These components jointly enable localized, adaptive reasoning and enhance the model's generalization across diverse meteorological conditions.

Our key contributions are as follows:

- **Pseudo-spectral inspired spatial encoding for flexible representation:** We propose a spatial transformation inspired by spectral methods that enables localized, non-periodic encoding on uniform grids, improving the model's ability to capture anisotropic and non-stationary precipitation patterns.
- **Graph-structured sparse token interaction with channel sparsification:** We develop a two-stage sparse modeling approach combining spatial clustering via graph construction and intra-cluster message passing, together with learnable channel sparsity. This design supports efficient, long-range spatial reasoning with enhanced interpretability.
- **Robust empirical performance across meteorological regimes:** Evaluated on ECMWF ensemble forecasts and TRMM daily precipitation data, our model consistently surpasses strong baselines across multiple geographic and seasonal contexts, demonstrating superior accuracy and generalization. Core code is included in the supplementary materials, with full code to be released upon acceptance.

2 Related Work

2.0.1 Spectral and Pseudo-Spectral Methods for Spatial Modeling.

Spectral methods, particularly the Fourier Neural Operator (FNO) Li et al. (2020), have shown promise in modeling spatially extended systems by capturing global interactions in the frequency domain. Extensions such as AFNO Guibas et al. (2021) introduce learnable filtering and spectral sparsification to better adapt to structured meteorological fields. However, their dependence on periodic boundary assumptions and global stationarity can limit applicability in tasks like precipitation diagnosis, which often involve sharp gradients, intermittent structures, and localized anomalies. To alleviate these issues, pseudo-spectral variants Zhao et al. (2020); Liu et al. (2024) replace global Fourier bases with more flexible alternatives such as DCT or data-driven transforms. Our approach builds on this idea by applying a learned spatial projection inspired by pseudo-spectral methods, which maintains grid alignment while enabling flexible spatial grouping.

2.0.2 Attention and Spatial Structuring via Clustering.

Transformer-based models have been widely adopted in weather prediction and geospatial analysis due to their ability to capture long-range dependencies. Windowed or clustered attention Kitaev et al. (2020); Roy et al. (2021); Zhu et al. (2023), as in the Swin Transformer Liu et al. (2021), mitigates

the computational cost of full attention. Our model leverages a domain-informed, frequency-guided clustering mechanism followed by graph-constrained attention within clusters. This structured approach encourages physically meaningful grouping while maintaining global modeling capacity through sparse inter-cluster communication and learnable channel-wise selection, offering a principled compromise between expressivity and efficiency.

2.0.3 PRECIPITATION DIAGNOSIS MODELS.

Recent work has explored precipitation diagnosis as a standalone task, separate from direct precipitation forecasting. On one end of the spectrum, FourCastNet Pathak et al. (2022) includes a lightweight diagnostic head that maps predicted state variables (such as humidity and temperature) to surface precipitation. While efficient and trainable end-to-end, this module applies a spatially uniform mapping and does not explicitly account for nonstationarity or multi-scale spatial patterns.

On the other end, Kochkov et al. (2024) propose a physically guided diagnostic approach based on vertically integrated moistening tendencies, where precipitation is derived from model-predicted constituent tendencies using simplified thermodynamic equations. This method benefits from physical interpretability, but is constrained by strong assumptions on vertical processes and lacks spatial adaptivity.

Our work offers an intermediate design: we introduce a spatial diagnostic framework that learns from data while preserving structural priors. By integrating structure-guided local encoding, cluster-wise attention, and spatial sparsification, our model adaptively captures local features, scale transitions, and spatial intermittency, providing a more expressive and interpretable alternative to existing rule-based or uniform neural diagnostic methods.

3 METHOD

3.1 PROBLEM FORMULATION

We consider the task of spatial precipitation diagnostics based on ensemble forecasts. Let t_0 denote the forecast initialization time for a given day D. The objective is to estimate the 24-hour accumulated precipitation field over day D, denoted by $\mathbf{Y} \in \mathbb{R}^{H \times W}$, where H and W correspond to the latitude and longitude dimensions of the forecast grid.

Instead of predicting the accumulated precipitation directly from the initial condition, we follow a diagnostic formulation wherein the target field is decomposed into the sum of instantaneous contributions derived from short-range forecasts. Specifically, we assume access to S consecutive forecast steps initialized at t_0, each with a lead time $\tau_s \in \{\tau_1, \ldots, \tau_S\}$, and each representing a temporal window $[t_s^{\text{start}}, t_s^{\text{end}}]$ such that $\bigcup_{s=1}^{S} [t_s^{\text{start}}, t_s^{\text{end}}] = D$.

For each forecast step $s \in \{1, \ldots, S\}$, the input tensor is given by $\mathbf{X}^{(s)} \in \mathbb{R}^{V \times M \times H \times W}$,where V denotes the number of prognostic variables, including surface and upper-air fields at multiple pressure levels, M is the number of ensemble members, $H \times W$ are the spatial dimensions.

Each step-wise input $\mathbf{X}^{(s)}$ corresponds to the instantaneous atmospheric state at time $t_0 + \tau_s$, and is used to infer the corresponding interval precipitation:

$$\widehat{\mathbf{Y}}^{(s)} = f_\theta \left(\mathbf{X}^{(s)} \right) \in \mathbb{R}^{H \times W}, \tag{1}$$

where f_θ denotes the diagnostic model parameterized by θ. The final prediction of daily accumulated precipitation is obtained by summing the step-wise estimates:

$$\widehat{\mathbf{Y}} = \sum_{s=1}^{S} \widehat{\mathbf{Y}}^{(s)}. \tag{2}$$

This formulation offers multiple advantages: (i) it leverages instantaneous meteorological information without resorting to autoregressive or recurrent prediction schemes, thus mitigating temporal error propagation; (ii) it enables model generalization across varying lead times; and (iii) it aligns with standard operational forecast formats.

In the primary case considered in this study, we utilize the ECMWF ENS dataset, where forecast outputs are available at 12-hour intervals. Accordingly, for predicting 24-hour precipitation, we adopt $S = 2$, corresponding to lead times of 12 and 24 hours. The full model input is then: $\mathbf{X} = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right\} \in \mathbb{R}^{S\times V\times M\times H\times W}$, with the aggregated diagnostic prediction:$\widehat{\mathbf{Y}} = f_\theta\left(\mathbf{X}^{(1)}\right) + f_\theta\left(\mathbf{X}^{(2)}\right)$.

This diagnostic setting forms the foundation of the model architecture described in the subsequent sections.

3.2 Model Overview

Our precipitation diagnostic model adopts an encoder-decoder architecture designed to process each forecast step independently and generate step-wise precipitation estimates, which are subsequently accumulated to yield daily totals.

For each step s, the model processes the ensemble forecast $\mathbf{X}^{(s)} \in \mathbb{R}^{V\times M\times H\times W}$ through two modules: an encoder ϕ that extracts hierarchical spatial representations, and a decoder ψ that produces the corresponding precipitation field. See the Appendix for the model structure diagram and pseudocode.

Specifically, the encoder ϕ maps the input into a set of multi-scale latent feature maps:

$$\left\{\mathbf{Z}_i^{(s)} \in \mathbb{R}^{C_i\times H_i\times W_i}\right\}_{i=1}^{L} = \phi\left(\mathbf{X}^{(s)}\right), \tag{3}$$

where i indexes the resolution level, with C_i denoting the number of channels and $H_i \times W_i$ the spatial resolution at scale i. The sequence $\{(C_i, H_i, W_i)\}_{i=1}^{L}$ defines a spatial pyramid, with coarser scales capturing broad structural patterns and finer scales retaining localized details.

The decoder ψ then reconstructs the step-wise precipitation field from the set of latent features:

$$\widehat{\mathbf{Y}}^{(s)} = \psi\left(\left\{\mathbf{Z}_i^{(s)}\right\}_{i=1}^{L}\right). \tag{4}$$

Finally, daily accumulated precipitation is obtained by summing over all forecast steps. This modular design enables flexible and interpretable modeling, where each forecast step is handled individually.

3.3 Encoder

The encoder ϕ is designed to capture multi-scale, anisotropic spatial patterns and correct potential biases in numerical precipitation forecasts through a structured decomposition and aggregation pipeline. The input at each forecast step $\mathbf{X}^{(s)} \in \mathbb{R}^{V\times M\times H\times W}$ consists of multi-variable, multi-member atmospheric states. We process it through four major stages: structure-guided local transformation, cluster-wise attention, graph-based routing, and spatial sparsification.

3.3.1 Localized Structural Encoding

To effectively capture non-stationary and anisotropic spatial structures associated with different precipitation regimes, we apply a localized structural encoding to each variable and ensemble member. Rather than relying on global assumptions such as stationarity or periodicity, we decompose the input fields into overlapping spatial patches, allowing the model to extract region-dependent variation patterns at multiple spatial scales.

This local treatment enables the model to distinguish between fine-scale, sharply varying convective systems and broader, smoothly varying stratiform patterns. Such spatial heterogeneity is often underrepresented in raw forecasts due to grid discretization and numerical diffusion. Our design facilitates scale-sensitive pattern recognition without relying on predefined analytic transforms, serving as a flexible preprocessing stage for downstream clustering and attention modules. To enhance region-specific structural expressiveness, our encoder applies a localized structural transformation to each spatial segment:

$$\widetilde{\mathbf{X}}^{(s)} = \mathcal{T}\left(\mathbf{X}^{(s)}\right), \tag{5}$$

where $\mathcal{T}$ denotes a learnable localized transformation that emphasizes salient spatial variations within each patch. The resulting representation $\widetilde{\mathbf{X}}^{(s)} \in \mathbb{R}^{V \times M \times H \times W \times F}$ captures enriched multi-scale features across channels, facilitating adaptive discrimination between convective and stratiform regimes. This step improves the model's ability to detect diagnostically relevant patterns that are often underrepresented in coarse-resolution forecasts.

3.3.2 Cluster-wise Spatial Encoding

Given $\widetilde{\mathbf{X}}^{(s)}$, we perform patch-level clustering in the spatial domain to form a set of localized attention units. The clustering is guided by frequency-enhanced embeddings, allowing patches with similar spectral profiles to be grouped together. Each patch p is assigned to a cluster $c(p)$, yielding K clusters across the domain. We denote the clustered input as:

$$\left\{\widetilde{\mathbf{X}}_c^{(s)}\right\}_{c=1}^{K}. \tag{6}$$

This cluster-wise structure enables efficient encoding of region-specific dynamics while preserving high-resolution variability.

3.3.3 Graph-constrained Cluster Attention

To preserve fine-grained dependencies within each cluster, we perform localized token fusion via soft aggregation, enabling the model to summarize intra-cluster spatial-frequency structures. Although our primary attention mechanism operates across clusters, this hierarchical design implicitly encodes intra-cluster interactions, further enhanced by the scale-sensitive cues captured via localized pseudo-spectral features and the efficiency-aware channel gating. We construct a sparse graph $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$ for each cluster c, where $\mathcal{V}_c$ corresponds to spatial patches and $\mathcal{E}_c$ encodes learnable connectivity informed by both spatial proximity and spectral similarity. We apply self-attention over each graph using shared parameters:

$$\mathbf{H}_c = \mathrm{GAttn}\left(\widetilde{\mathbf{X}}_c^{(s)}; \mathcal{G}_c\right), \tag{7}$$

producing graph-aggregated representations $\mathbf{H}_c$ for each cluster. The constrained attention mechanism promotes physical interpretability and limits spurious interactions, especially in synoptic-scale disturbances.

3.3.4 Channel Sparsification and Fusion

To enhance efficiency and reduce redundancy, we introduce a learnable channel-wise sparsity mechanism. Specifically, for each cluster-level representation $\mathbf{H}_c \in \mathbb{R}^{C \times H_c \times W_c}$, we apply a soft-thresholding operator $\Gamma(\cdot)$ to suppress low-energy or redundant channel activations. The function $\Gamma(\cdot)$ is defined as:

$$\Gamma(x_i; \lambda_i) = \mathrm{sign}(x_i) \cdot \max\left(|x_i| - \lambda_i, 0\right), \quad \forall i \in [1, C] \tag{8}$$

where x_i denotes the activation in the i-th channel, and λ_i is a learnable non-negative threshold specific to each channel.

This operation encourages sparsity in the latent representation, promoting compact and efficient coding. By selectively retaining informative frequency channels, the model enhances both generalization and interpretability, as retained channels tend to capture localized spectral patterns relevant to regional precipitation variability.

3.4 Decoder

The decoder reconstructs the final prediction from multi-scale encoder features $\mathcal{Z} = \left\{\mathbf{Z}_l \in \mathbb{R}^{C_l \times H_l \times W_l}\right\}_{l=0}^{L-1}$, where L is the number of scales. Optionally, it incorporates static geographical information $\mathbf{E} \in \mathbb{R}^{D_s \times H \times W}$ defined on the target grid of size $H \times W$.

To facilitate fusion, each encoder feature $\mathbf{Z}_l$ is first bilinearly interpolated to the target resolution:

$$\widetilde{\mathbf{Z}}_l = \mathrm{Interp}_{\text{bilinear}}(\mathbf{Z}_l; H, W). \tag{9}$$

If static features are available ($D_s > 0$), they are similarly resized:

$$\widetilde{\mathbf{E}} = \text{Interp}_{\text{bilinear}}(\mathbf{E}; H, W). \tag{10}$$

These are concatenated along the channel dimension and fused by a 1×1 convolution:

$$\mathbf{Z}_l^{\text{fused}} = \mathbf{W}_l^m * \left[\widetilde{\mathbf{Z}}_l; \widetilde{\mathbf{E}}\right] + \mathbf{b}_l^m, \tag{11}$$

where $\mathbf{W}_l^m \in \mathbb{R}^{C_l \times (C_l + D_s)}$ and $\mathbf{b}_l^m$ are learnable parameters. This stage aligns heterogeneous features and incorporates domain knowledge into dynamic representations.

To enhance spectral coherence and suppress aliasing artifacts, a frequency-aware transformation is introduced. Let $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the 2D Discrete Fourier Transform (DFT) and its inverse, respectively. A learnable modulation kernel

$$\mathbf{K}_l^{\text{spec}} \in \mathbb{C}^{H \times W \times C_l} \tag{12}$$

is applied element-wise in the frequency domain:

$$\mathbf{Z}_l^{\text{freq}} = \mathcal{F}^{-1}\left(\mathbf{K}_l^{\text{spec}} \odot \mathcal{F}(\mathbf{Z}_l^{\text{fused}})\right). \tag{13}$$

The resulting feature undergoes spatial-domain refinement to restore fine-scale details:

$$\mathbf{Z}_l^{\text{refined}} = \text{Conv}_l^{3\times 3}\left(\sigma\left(\text{Conv}_l^{3\times 3}(\mathbf{Z}_l^{\text{freq}})\right)\right) + \mathbf{Z}_l^{\text{freq}}, \tag{14}$$

where $\sigma(\cdot)$ denotes the GELU activation. The residual structure preserves high-frequency content while reducing noise.

Each scale's refined feature is projected into a shared latent space of dimension F via a 1×1 convolution:

$$\mathbf{Z}_l^{\text{aligned}} = \mathbf{W}_l^a * \mathbf{Z}_l^{\text{refined}} + \mathbf{b}_l^a, \tag{15}$$

where $\mathbf{W}_l^a \in \mathbb{R}^{F \times C_l}$ and $\mathbf{b}_l^a$ are learnable parameters.

To fuse multi-scale information, a scale-wise attention mechanism is used. First, global average pooling is applied to obtain summary descriptors:

$$\mathbf{g}_l = \text{GAP}(\mathbf{Z}_l^{\text{aligned}}). \tag{16}$$

These are passed through a small MLP to generate attention logits:

$$s_l = \mathbf{W}^2 \cdot \sigma(\mathbf{W}^1 \cdot \mathbf{g}_l + \mathbf{b}^1) + \mathbf{b}^2. \tag{17}$$

The attention weights are normalized via softmax:

$$\alpha_l = \frac{\exp(s_l)}{\sum_{k=0}^{L-1} \exp(s_k)}. \tag{18}$$

The fused representation aggregates aligned features across scales:

$$\mathbf{Z}^{\text{fused}} = \sum_{l=0}^{L-1} \alpha_l \cdot \mathbf{Z}_l^{\text{aligned}}. \tag{19}$$

Finally, the output is predicted via a two-stage convolutional head:

$$\mathbf{Y} = \mathbf{W}^{\text{out}} * \sigma\left(\mathbf{W}^h * \mathbf{Z}^{\text{fused}} + \mathbf{b}^h\right) + \mathbf{b}^{\text{out}}, \tag{20}$$

where $\mathbf{W}^h \in \mathbb{R}^{F \times F}$ and $\mathbf{W}^{\text{out}} \in \mathbb{R}^{C_{\text{out}} \times F}$ are implemented as 3×3 and 1×1 convolutions, respectively. The output $\mathbf{Y} \in \mathbb{R}^{C_{\text{out}} \times H \times W}$ corresponds to the final spatial prediction (e.g., precipitation intensity when $C_{\text{out}} = 1$).

This decoder design integrates dynamic and static information via multi-scale alignment and fusion. Frequency-aware upsampling mitigates interpolation artifacts and sharpens spatial structure. Meanwhile, the attention-based fusion promotes adaptive weighting across scales, improving generalization and resolution consistency.

4 EXPERIMENTS

4.1 DATASETS

Our experiments utilize ensemble forecast data (ENS) from 2010 to 2018, alongside the TEMM daily precipitation observations. Due to data availability constraints, some months are missing; a detailed list of missing periods is provided in the Appendix.

The ENS forecasts include five upper-air variables measured at four standard pressure levels, as well as seven surface variables. The specific variables and pressure levels are enumerated in the Appendix.

It is important to note a slight spatial discrepancy between the ENS and TRMM domains. The latitude and longitude grids are defined as follows:

$$\begin{aligned}
\texttt{lats_ENS} &= [35.0, 34.75, \ldots, 10.0], \\
\texttt{lons_ENS} &= [100.0, 100.25, \ldots, 125.0], \\
\texttt{lats_TRMM} &= [35.125, 34.875, \ldots, 10.125], \\
\texttt{lons_TRMM} &= [99.875, 100.125, \ldots, 124.875].
\end{aligned}$$

For models incorporating a positional encoding module, we preprocess the ENS input fields by bilinearly interpolating them onto the TRMM grid. The TRMM grid coordinates are then used to generate positional encodings. This procedure ensures consistent spatial reference frames and improves the positional information quality.

For models without explicit positional encoding, we do not apply interpolation to avoid introducing additional interpolation-induced biases into the ENS forecast data. We assume that these models implicitly learn to handle the slight spatial offset during training.

All input forecast variables are normalized independently using their climate statistics computed per variable and pressure level. This normalization mitigates scale differences across inputs and facilitates stable training.

4.2 EVALUATION

To assess the effectiveness of our method, we compare it against a suite of representative baselines commonly used in spatial precipitation prediction tasks:

- **LeNet** Li et al. (2022): A simple convolutional neural network, originally proposed for handwritten digit classification, adapted here for multivariate bias correction and single-variable regression. Its lightweight structure makes it a reasonable lower-bound baseline.
- **U-Net + WMSE-MS-SSIM loss** Hess & Boers (2022): A modified U-Net structure designed for precipitation post-processing. In addition to weighted MSE (WMSE), it incorporates a multi-scale Structural Similarity Index (MS-SSIM) in the loss function to enhance sensitivity to local structures and spatial intermittency in precipitation.
- **FourCastNet (precipitation head)** Pathak et al. (2022): We employ the diagnostic branch of FourCastNet, which follows an Adaptive Normalization-based Fourier Operator (ANFO) architecture. This model represents the use of frequency-domain modeling in high-resolution precipitation diagnostics.
- **Swin-Unet** Cao et al. (2022): A hierarchical Transformer-based architecture employing shifted window self-attention. It has demonstrated competitive performance in learning the spatial patterns of meteorological variables Li et al. (2025), serving as a strong general-purpose spatial model.

We evaluate model performance using both continuous and categorical metrics.

For point-wise numerical evaluation, we report three widely used regression metrics: Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Pearson Correlation Coefficient (CORR). See the Appendix for detailed formulas.

Due to the intermittent and threshold-sensitive nature of precipitation, we additionally evaluate models using two categorical skill scores based on event detection:

- **Equitable Threat Score (ETS)**: The ETS is computed for each rainfall threshold T_k from the predefined set $1.0, 10.0, 25.0, 50.0, 100.0$ mm/day, which correspond to [light, moderate, heavy, rainstorm, severe rainstorm]. For each threshold, we binarize the fields:

$$\hat{B}_i^{(k)} = \mathbb{I}[\hat{P}_i \geq T_k], \quad B_i^{(k)} = \mathbb{I}[P_i \geq T_k]. \tag{21}$$

  Then the contingency table components (a_k, b_k, c_k, d_k) are: a_k: hits (both predict and observe above threshold), b_k: false alarms (predict above, observe below), c_k: misses (predict below, observe above), d_k: correct rejections.
  The ETS at level k is:

$$\text{ETS}_k = \frac{a_k - a_k^{\text{rand}}}{a_k + b_k + c_k - a_k^{\text{rand}}}, \tag{22}$$

  where $\quad a_k^{\text{rand}} = \frac{(a_k+b_k)(a_k+c_k)}{a_k+b_k+c_k+d_k}$

- **Hanssen–Kuipers (HK) score**: We adopt a multi-class extension of the Hanssen–Kuipers score (also known as the True Skill Statistic, TSS), by averaging class-wise TSS across all categories. For each category i, the score is computed as:

$$\text{TSS}_i = \frac{C_{ii}}{\sum_j C_{ij} + \epsilon} - \frac{\sum_{j \neq i} C_{ji}}{\sum_{j \neq i} \sum_k C_{jk} + \epsilon} \tag{23}$$

  where C_{ij} denotes the number of grid cells predicted as category i and observed as category j. The final HK score is obtained by averaging over all K categories:

$$\text{HK} = \frac{1}{K} \sum_{i=1}^{K} \text{TSS}_i \tag{24}$$

4.3 Training

All models were trained with a batch size of 24 for up to 30 epochs. We used the AdamW optimizer with an initial learning rate of 10^{-4}. A linear warmup strategy was applied during the first epoch. For most models, we adopted a Reduce-on-Plateau scheduler with a decay factor of 0.1 and patience of 2 epochs. The only exception is FourCastNet, for which we followed the original implementation and used a cosine annealing schedule down to 10^{-5}.

We applied an early stopping strategy: training was terminated if the validation loss did not improve for 4 consecutive epochs.

Training was conducted on NVIDIA H100 GPUs, with a maximum of 3 GPUs utilized per model. Lightweight models such as LetNet and UNet were trained on 2 GPUs. The full training set comprises 5170 samples, of which 3712 from 2010–2016 were used for training, and 1458 from 2017–2018 were used for both validation and testing. During testing, we set the batch size to 1 and evaluated models on a single GPU.

The model size, total training and validation time, and GPU usage are summarized in Appendix. We can see that although our model has the most parameters, it does not consume more resources (compared to FourCastNet and SwinUnet).

5 Results

5.1 Quantitative Evaluation

We begin the evaluation by comparing the overall performance of baseline models and our proposed method on the test set, using both general-purpose regression metrics and precipitation-specific classification scores. The results are shown in Table 1 and Table 2.

Across all four general-purpose metrics—Root Mean Square Error (RMSE), Coefficient of Determination (R^2), Pearson Correlation Coefficient (Corr), and Hanssen–Kuipers (HK) score—our model

Model	RMSE ↓	R^2 ↑	Corr ↑	HK ↑
LeNet	10.83686	0.17647	0.55321	0.07936
UNet	10.99117	0.12465	0.52933	0.18096
SwinUnet	10.53069	0.23028	0.56670	0.11775
FourCastNet	10.44083	0.23944	0.56513	0.15167
Ours	**9.43784**	**0.37276**	**0.61705**	**0.20059**

Table 1: Overall performance on the test set in terms of RMSE (↓), R^2 (↑), Corr (↑), and HK (↑). Lower RMSE and higher values for the other metrics indicate better performance.

outperforms the baselines by a clear margin. In particular, it improves RMSE by more than 1.0 mm/6hr compared to the next best method (FourCastNet), while boosting R^2 from 0.239 to 0.373 (+13.3%) and Corr from 0.565 to 0.617 (+5.2%). These results indicate that our model not only reduces absolute prediction errors, but also strengthens its ability to reconstruct the spatial variability and linear relationship with the ground truth precipitation.

The HK score, which reflects multi-class classification skill (using a multi-class extension), further validates this improvement from a categorical prediction standpoint. Compared to FourCastNet, our method achieves a +4.5% gain in HK, suggesting that the improvement is consistent across different rainfall categories and not only for average or dominant regimes.

To investigate performance across precipitation intensities, Table 2 reports the Equitable Threat Score (ETS) under five rainfall thresholds $[1.0, 10.0, 25.0, 50.0, 100.0]$ mm/24hr, corresponding to light rain, moderate rain, heavy rain, rainstorm, and severe rainstorm, respectively. Unlike HK, ETS is a binary metric computed separately at each threshold, and is particularly sensitive to both hit rate and false alarm rate for each precipitation category.

Our model achieves the best ETS scores at all thresholds above 10.0 mm, i.e., moderate to extreme rainfall levels. In particular, ETS(25.0) and ETS(50.0) improve by +2.4% and +3.3%, respectively, compared to the UNet baseline, and ETS(100.0) improves dramatically from 0.00888 (UNet) to 0.07194. This confirms the model's strength in detecting rare but high-impact events—an essential trait in real-world forecasting and decision-making applications.

However, the ETS(1.0) value for our model is slightly lower than some baselines, indicating a trade-off in performance for light precipitation events. This is expected, as light rain is often more spatially fragmented and less predictable, and optimizing for high-intensity accuracy tends to under-represent low-end events. Nevertheless, the increase in HK—being a multi-class skill score—suggests that the overall classification capability remains strong and well-balanced.

Taken together, RMSE and R^2 capture absolute error and explained variance, indicating that our model provides more accurate and reliable spatial precipitation estimates. Corr measures the linear correspondence between predicted and observed patterns, showing that our method preserves spatial gradients better than others. HK (multi-class) and ETS (binary, per class) jointly highlight the categorical accuracy. Our model maintains high accuracy for moderate rain and significantly outperforms all baselines in the heavy and extreme rainfall regimes—crucial for risk assessment.

The superior ETS at higher thresholds aligns well with the lower RMSE and higher R^2, implying that the model is not simply averaging out large events to minimize L2 loss, but is actively learning to recover their spatial extent and magnitude. These results collectively demonstrate that our method achieves a better balance between regression precision and categorical sharpness, and is particularly advantageous for applications where heavy rain detection is vital.

5.2 Spatial Distribution of Model Accuracy

To further investigate where different models perform well or poorly, we visualize the spatial distribution of three representative evaluation metrics: RMSE, Pearson correlation coefficient (CORR), and Hanssen–Kuipers (HK) score, as shown in Figure 1. These maps reflect the spatially averaged scores across the entire testing period, offering intuitive insight into each model's geographic strengths and weaknesses.

Model	1.0	10.0	25.0	50.0	100.0
LeNet	27.22	16.11	5.694	0.652	0.021
UNet	25.26	22.06	20.16	13.97	0.888
SwinUnet	27.87	17.87	9.833	4.259	0.000
FourCastNet	26.32	18.82	9.274	4.247	0.246
Ours	21.21	**28.87**	**22.56**	**17.27**	**7.194**

Table 2: Equitable Threat Score (ETS $\times 10^{-2}$) at different rainfall thresholds (1.0, 10.0, 25.0, 50.0, 100.0 mm/day) on the test set. Higher values indicate better categorical prediction performance at corresponding precipitation intensities.
* All ETS values are scaled by 10^2 for compact presentation. Original values can be obtained by dividing displayed numbers by 100.

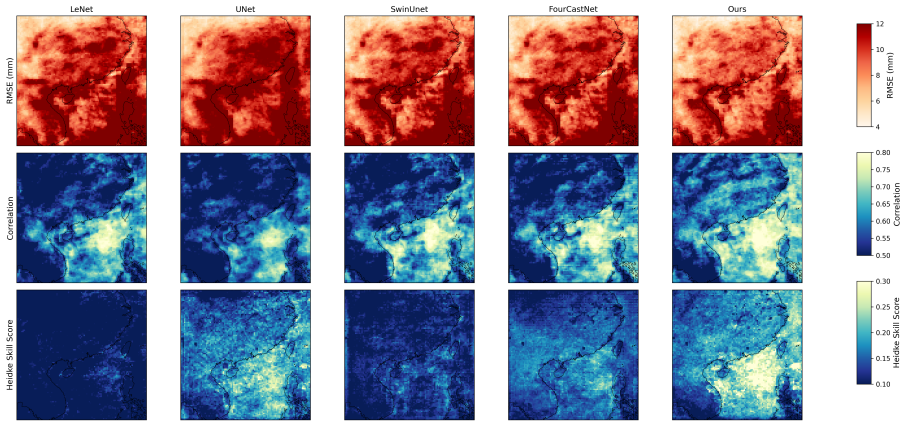

Figure 1: Spatial distribution of RMSE, CORR and HK.

The spatial patterns of model performance exhibit notable consistency across all models. For RMSE, prediction errors are substantially higher over oceanic regions, with particularly pronounced degradation along coastal areas—most notably the eastern coastline of Southeast Asia. Despite this general trend, our model consistently outperforms baselines across most regions and exhibits a narrower ”band of degradation,” indicating better generalization in challenging coastal environments.

In contrast, the correlation coefficient (CORR) shows an opposite spatial tendency: higher values are concentrated near coastal zones, while inland areas suffer from reduced correlation. This pattern likely reflects the stronger deterministic signals near oceans and the greater stochastic variability over land. Again, our model maintains relatively stable performance, with smaller drops in CORR compared to others, although inland degradation is still present.

The spatial distribution of the Hanssen–Kuipers (HK) score is notably more uniform across regions, suggesting that the categorical detection skill of all models is less sensitive to geographic context. However, both our model and UNet stand out with slightly better HK scores near coastal regions—mirroring the CORR patterns—while other models show little regional variation. Considering that these two models also rank highest in overall HK score, it highlights our model's ability to retain detection skill where other models begin to fail. This reinforces the conclusion that our model not only achieves higher quantitative performance but also demonstrates more spatially robust predictive skill, especially in transitional zones like coastal boundaries.

5.3 Robustness and Stability

We visualize the full distribution of model performance across the test set using boxplots of HK, ETS25, ETS50, and ETS100 scores (Figure 2). Overall, our model demonstrates significantly stronger and more stable performance compared to all baselines on most metrics. In particular, for HK, our model achieves notably higher median values, with several individual cases exceeding

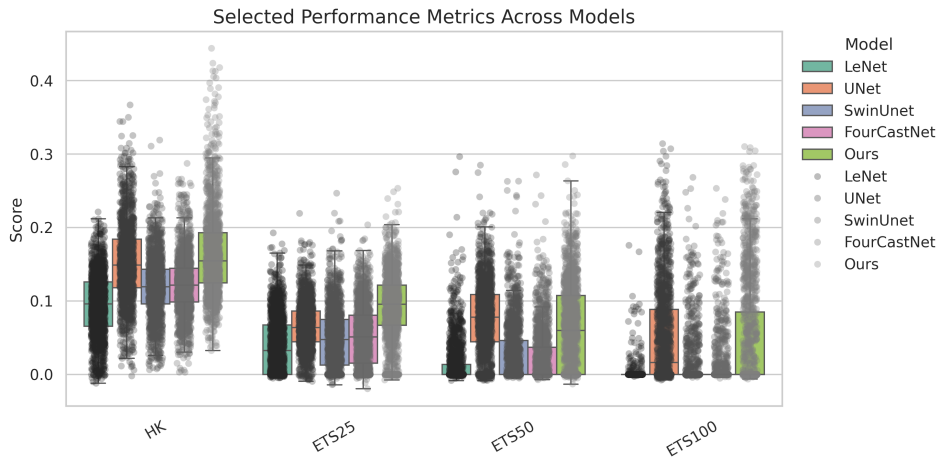

Figure 2: Boxplot of ETS and HK distribution between samples.

an HK of 0.4—a level rarely approached by other models. This suggests that our method not only performs well on average but also maintains robustness across diverse testing scenarios.

Interestingly, our model also performs strongly under high-threshold precipitation conditions. While the performance gap on ETS50 is narrower than for other thresholds, our model still achieves comparable or better scores in extreme rainfall cases. This indicates that the model is capable of capturing both general patterns and rare but critical extreme events.

Among the baselines, the UNet variant trained with SSIM supervision stands out, delivering second-best results on all evaluated metrics. We attribute this to the perceptual sensitivity of the SSIM loss, which may help better preserve local structural features in the precipitation field. In contrast, our model is trained purely with MSE loss, yet it outperforms or matches SSIM-UNet in both central tendency and spread of metrics, underscoring the effectiveness of our architectural design in learning spatial structures without relying on perceptual losses.

Overall, the small interquartile ranges and high median values across HK and ETS metrics reflect not only the accuracy but also the stability of our model, especially in comparison with other spatially aware baselines.

6 Conclusion

In this work, we proposed a novel spatial diagnostic model tailored for high-resolution precipitation prediction. Unlike traditional methods that rely heavily on temporal dynamics, our approach is designed to operate solely in the spatial domain, addressing a critical need in spatial-based postprocessing and interpretation tasks. The model introduces a structured pipeline: a pseudo spectral-space transformation to highlight spatial modes, followed by cluster-wise attention that enables localized, interpretable feature fusion, and a sparse channel gating mechanism to enhance model compactness and robustness. The architecture is further designed to support recovery to the original spectral-temporal space with minimal information loss.

Through extensive experiments, our model demonstrates consistent and superior performance across multiple evaluation metrics, especially under challenging conditions such as extreme rainfall events. It outperforms several strong baselines, including those incorporating perceptual losses like SSIM, despite being trained with a simple MSE objective. The proposed method not only achieves higher median scores but also exhibits lower variance, highlighting both its accuracy and robustness across diverse samples.

Future work could explore integrating temporal information, expanding to multivariate spatial fields, or applying the architecture to other geophysical variables with complex spatial intermittency.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. To this end, we provide the following:

Code and Model Implementation: All code used for experiments, including model architectures, training scripts, and evaluation metrics, will be released under a permissive open-source license (MIT License) upon publication. The repository includes detailed documentation and instructions for reproducing the reported results.

Data Access: The datasets used in this study are either publicly available (e.g., ENS Swinbank et al. (2016) and TRMM Liu et al. (2012)).

Hyperparameters and Training Details: We report all critical hyperparameters in the paper (e.g., learning rate, batch size, optimizer settings, network architecture). Full training configurations (including random seeds, augmentation policies, and convergence criteria) are included in the supplementary material and code repository.

Experimental Setup: All experiments were conducted using PyTorch [3] on NVIDIA H100 GPUs.

We believe this information is sufficient for other researchers to fully reproduce our results and build upon our work.

REFERENCES

Pierre Bourgouin. A method to determine precipitation types. *Weather and Forecasting*, 15(5): 583–592, 2000.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

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.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Philipp Hess and Niklas Boers. Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, 14(3):e2021MS002765, 2022.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

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.

René Laprise. Regional climate modelling. *Journal of computational physics*, 227(7):3641–3666, 2008.

Wentao Li, Baoxiang Pan, Jiangjiang Xia, and Qingyun Duan. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of hydrology*, 605:127301, 2022.

Zhuang Li, Zhenyu Lu, Yizhe Li, and Xuan Liu. Swinnowcast: A swin transformer-based model for radar-based precipitation nowcasting. *Remote Sensing*, 17(9):1550, 2025.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Huiwen Liu, Ying-Bo Lu, Li Zhang, Fangchao Liu, You Tian, Hailong Du, Junsheng Yao, Zi Yu, Duyi Li, and Xuemai Lin. Pseudo-spectral spatial feature extraction and enhanced fusion image for efficient meter-sized lunar impact crater automatic detection in digital orthophoto map. *Sensors*, 24(16):5206, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zhong Liu, Dana Ostrenga, William Teng, and Steven Kempler. Tropical rainfall measuring mission (trmm) precipitation data and services for research and applications. *Bulletin of the American Meteorological Society*, 93(9):1317 – 1325, 2012. doi: 10.1175/BAMS-D-11-00152.1. URL https://journals.ametsoc.org/view/journals/bams/93/9/bams-d-11-00152.1.xml.

Miguel Nogueira. *The dynamics of convective orographic precipitation: the transition to convection*. PhD thesis, Universidade de Lisboa (Portugal), 2014.

Young-Youn Park, Roberto Buizza, and Martin Leutbecher. Tigge: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(637):2029–2050, 2008.

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.

Christopher D Roberts, Retish Senan, Franco Molteni, Souhail Boussetta, Michael Mayer, and Sarah PE Keeley. Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Geoscientific model development*, 11(9):3681–3712, 2018.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.

Joanne Simpson, C Kummerow, W-K Tao, and Robert F Adler. On the tropical rainfall measuring mission (trmm). *Meteorology and Atmospheric physics*, 60(1):19–36, 1996.

Richard Swinbank, Masayuki Kyouda, Piers Buchanan, Lizzie Froude, Thomas M. Hamill, Tim D. Hewson, Julia H. Keller, Mio Matsueda, John Methven, Florian Pappenberger, Michael Scheuerer, Helen A. Titley, Laurence Wilson, and Munehiko Yamaguchi. The tigge project and its achievements. *Bulletin of the American Meteorological Society*, 97(1):49 – 67, 2016. doi: 10.1175/BAMS-D-13-00191.1. URL https://journals.ametsoc.org/view/journals/bams/97/1/bams-d-13-00191.1.xml.

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.

Xuebin Zhao, Hui Zhou, Hanming Chen, and Yufeng Wang. Domain decomposition for large-scale viscoacoustic wave simulation using localized pseudo-spectral method. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2666–2679, 2020.

Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10323–10333, 2023.

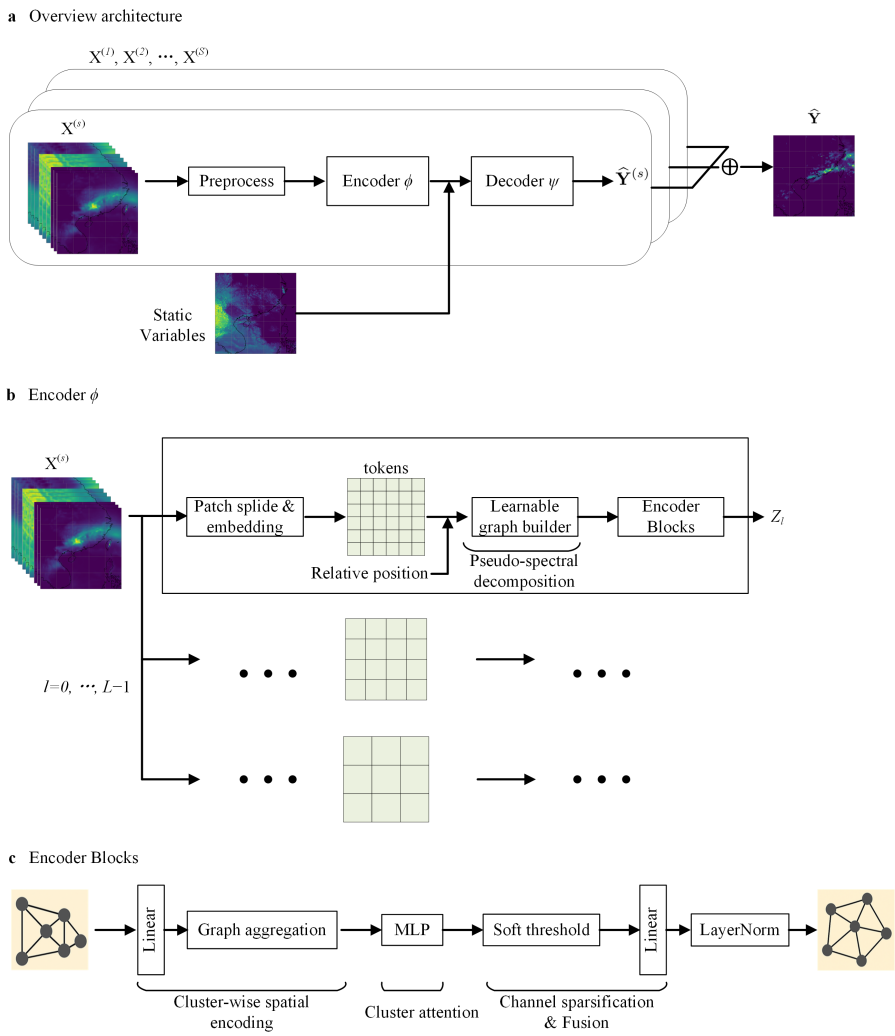

Figure 3: An overview diagram of the proposed neural network.

A Appendix

B A: Method

B.1 A1: Network Architecture

See Figure 3. a is the overall model architecture. b is the overall structure of the encoder. c is the encoder core layer structure.

[tb] **Input**: Token features $\mathbf{T} \in \mathbb{R}^{B \times N \times D}$; spatial coordinates $\mathbf{P} \in \mathbb{R}^{N \times 2}$
Output: Sparse edge index $\mathbf{E} \in \mathbb{N}^{2 \times E}$; edge weights $\mathbf{A} \in \mathbb{R}^{B \times E}$ [1] Fit k-NN on $\mathbf{P}$ to get top-$(k{+}1)$ indices $\mathbf{I} \in \mathbb{N}^{N \times (k+1)}$ Remove self-loop by discarding $\mathbf{I}_{:,0}$ Let $\texttt{neighbors} \leftarrow \mathbf{I}_{:,1:k+1}$ Set $\texttt{row} \leftarrow \texttt{repeat}(\texttt{arange}(N), k)$ Set $\texttt{col} \leftarrow \texttt{neighbors.reshape}(-1)$ Form edge index: $\mathbf{E} = \begin{bmatrix} \texttt{row} \\ \texttt{col} \end{bmatrix} \in \mathbb{N}^{2 \times E}$ Extract relative coordinates: $\mathbf{P}_i = \mathbf{P}[\mathbf{E}_0]$, $\mathbf{P}_j = \mathbf{P}[\mathbf{E}_1]$ Compute relative position encoding: $\texttt{rel_pos_emb} \leftarrow \texttt{RelPosEnc}(\mathbf{P}_i, \mathbf{P}_j) \in \mathbb{R}^{E \times D}$ Gather tokens: $\mathbf{T}_i = \mathbf{T}[:, \mathbf{E}_0, :]$, $\mathbf{T}_j = \mathbf{T}[:, \mathbf{E}_1, :]$ Concatenate edge features: $\mathbf{F}_{\text{edge}} = [\mathbf{T}_i, \mathbf{T}_j, \texttt{rel_pos_emb}] \in \mathbb{R}^{B \times E \times 3D}$ Compute edge logits: $\mathbf{L}_{\text{edge}} = \texttt{MLP}(\mathbf{F}_{\text{edge}})$ Normalize edge weights: $\mathbf{A} = \texttt{softmax}(\mathbf{L}_{\text{edge}}, \text{dim} = E)$ **return** $\mathbf{E}$, $\mathbf{A}$

B.2 A2: Pseudo-Code of Components

This algorithm implements a learnable graph construction module called the "Learnable Graph Builder" for localized structural encoding, as shown in Algorithm B.1. Given a batch of spatial token features and their 2D coordinates, it first performs a k-nearest neighbors (k-NN) search to identify each node's local neighbors, forming a preliminary graph structure. For each edge, relative positional encodings between connected nodes are computed and concatenated with their token features. This combined edge feature is then passed through a multilayer perceptron (MLP) to produce edge logits, which are normalized with a softmax to obtain dynamic, sparse edge weights. This approach enables adaptive, data-driven modeling of local spatial dependencies, enhancing the encoder's ability to capture complex spatial patterns while maintaining computational efficiency and interpretability.

As illustrated in Algorithm B.2, the graph aggregation module computes cluster-wise spatial encoding via message passing over the constructed graph. For each batch, the projected node features are aggregated from their graph neighbors using weighted summation, where the edge weights reflect affinity between clusters. This results in localized structural representations that are further processed by the downstream attention module, which is implemented as a simple MLP. [tb] **Input**: Token features $x \in \mathbb{R}^{B\times N\times C}$, graph edges $\texttt{edge_index} \in \mathbb{Z}^{2\times E}$, edge weights $\texttt{edge_weight} \in \mathbb{R}^{E}$
Output: Aggregated representation $x_{\text{agg}} \in \mathbb{R}^{B\times N\times H}$ [1] Project input tokens: $x \leftarrow \texttt{proj_in}(x)$ Initialize x_{agg} as zeros with shape (B, N, H) Extract source and target indices: $(s, t) \leftarrow \texttt{edge_index}$ each batch $b = 1$ to B Flatten node indices: $s_b \leftarrow s + b \cdot N$, $t_b \leftarrow t + b \cdot N$ Fetch destination features: $x_t \leftarrow x[t_b]$ Compute weighted messages: $m \leftarrow x_t \odot \texttt{edge_weight}$ Accumulate messages: $x_{\text{agg}}[s_b] \mathrel{+}= m$ **return** x_{agg}

C B: Experiments

C.1 B1: Data Validity Period

After data cleaning and temporal alignment, our dataset spans the years 2010–2018. However, due to missing or invalid entries, only a subset of months remains usable for training and evaluation. Specifically, valid data is available for the following months:

- 2010: February-December (missing January)
- 2011: March-December (missing October)
- 2012: January-October (missing February, May, June)
- 2013-2014: January-September (sporadic gaps)
- 2015: Complete coverage
- 2016:
- 2017-2018: Complete coverage

Importantly, the boreal summer rainy season (May to October) is largely preserved across these years, ensuring sufficient coverage of seasonal precipitation patterns critical for our task.

C.2 B2: Input Variables

Table 4 is a list of input variables.

C.3 B3: Formulas of Regression Metrics

- **Root Mean Square Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{25}$$

Model	Parameters (M)	Time (h)	GPUs
UNet	0.647	5.4	2
LeNet	0.055	9.9	2
SwinUnet	27.833	8.5	3
FourCastNet	106.786	13.1	3
Ours	149.964	8.7	3

Table 3: Training Resource Consumption

- **Coefficient of Determination** (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{26}$$

- **Pearson Correlation Coefficient (CORR)**:

$$\text{CORR} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}} \tag{27}$$

Here, $\hat{y}_i$ and y_i denote the predicted and observed values at grid point i, respectively; N is the number of spatial grid points or spatiotemporal samples, and $\bar{y}$, $\bar{\hat{y}}$ denote the sample means.

C.4 B4: Training Detials

The model size, total training and validation time, and GPU usage are summarized in Table 3. We can see that although our model has the most parameters, it does not consume more resources (compared to FourCastNet and SwinUnet).

D C: Results

Figure 1 show the spatial distribution of three representative evaluation metrics: RMSE, Pearson correlation coefficient (CORR), and Hanssen–Kuipers (HK) score.

Category	Variables (Description)
Surface (Single-Level)	`t2m` : 2m air temperature
	`skt` : skin temperature
	`d2m` : 2m dew point temperature
	`msl` : mean sea level pressure
	`u10` : 10m zonal wind
	`v10` : 10m meridional wind
	`sp` : surface pressure
	`tp` : total precipitation
5*Vertical (Multi-Level)	`u` : zonal wind (levels: 850, 700, 500, 250 hPa)
	`v` : meridional wind (same levels)
	`t` : air temperature (same levels)
	`gh` : geopotential height (same levels)
	`q` : specific humidity (same levels)
Static	`zs` : surface geopotential (orography)

Table 4: List of Input Variables