GENERATIVE HIERARCHICAL MESH NETWORK FOR GLOBAL STATION WEATHER PROBABILISTIC FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Global Station Weather Forecasting (GSWF) is pivotal for energy, aviation, and agriculture, yet existing time series methods struggle with its core challenges: capturing hierarchical dependencies and inherent uncertainty of chaotic meteorological systems, while adapting to global weather station data's uneven spatial distribution and limited size. To address these, we propose a novel Generative Hierarchical Mesh Network (GHMN). Specifically, we construct a hierarchical graph with bidirectional cross-level mapping, encoding bottom-layer local details into top-layer global features, and propagate global context downward for fine-grained predictions. For uncertainty modeling, we design a lightweight flow matching-based generative head that takes deterministic outputs as conditions to generate plausible ensemble predictions without preset priors. By decoupling forecasting into deterministic mean prediction and residual uncertainty estimation, we reduce learning complexity. Experiments show GHMN outperforms spatiotemporal baselines by up to 11%, with the advantage of probabilistic prediction. Our code is available at `https://anonymous.4open.science/r/GHMN-334C`.

1 INTRODUCTION

Global Station Weather Forecasting (GSWF) is vital to modern society, with significant implications for various sectors including energy (Dehalwar et al., 2016), aviation (Gultepe et al., 2019), and agriculture (Ukhurebor et al., 2022). Unlike regular image-like data structures such as Earth Reanalysis 5 (Hersbach et al., 2020), weather station data comprises precise, fine-grained meteorological observations distributed across irregular spatial locations. Currently, the most widely used and supposedly most accurate approach for GSWF is to apply Numerical Weather Prediction (NWP) models, which are developed by either solving physical equations or pretraining on structured grids, and then map the results to station locations via interpolation algorithms. However, physics-based NWP models are inherently computationally intensive, demanding extensive resources in execution. Most machine learning-based NWP models are trained and validated on reanalysis data, which integrates historical observations with simulated fields. Such integration may introduce systematic biases, particularly in regions with sparse observations or complex terrain (Li et al., 2025). While temporally consistent and spatially smooth, these data are not equivalent to real-world station observations; thus, these models may not be optimal for GSWF.

Recently, spatiotemporal series forecasting methods, which leverage historical weather observations and station spatial location information to predict future conditions, have exhibited superior performance and offer a more cost-effective alternative to traditional NWP models. Graph neural networks (GNNs) have proven effective in modeling spatiotemporal patterns across global stations (Ni et al., 2022; Hettige et al., 2024; Xu et al., 2024b; Tian et al., 2025; Zheng et al., 2025). However, these models typically only model spatial correlation based on spatial proximity, overlooking the global connectivity of the Earth's weather system, where weather patterns in one region are frequently influenced by conditions in distant regions. To sidestep this, Transformer-based models have gained widespread interest, as the attention mechanism enables instant communication between any two nodes and thereby effectively captures global spatial correlation (Wu et al., 2023; Liang et al., 2023; Yu et al., 2024; Chen et al., 2025). However, most existing methods perform forecasting at a single scale, hindering the capture of interactions between multi-scale processes of complex climate dynam-

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.

ics. In fact, the physical processes of the climate system unfold across different spatiotemporal scales, and their intrinsic complexity stems from cross-scale interactions (Reichstein et al., 2019; Bailie et al., 2025). For example, El Niño–La Niña cycles (McPhaden et al., 2006) drive large-scale ocean circulation while simultaneously shaping local-scale phenomena. Existing methods are often not specifically designed to capture and leverage these characteristics, ultimately limiting their predictive performance and generalization ability.

Furthermore, another key limitation of existing GSWF models is their general neglect of probabilistic forecasting. Most of these methods train models with deterministic loss functions, generating overly smoothed predictions that struggle to characterize extreme weather conditions (Sun et al., 2024) and fail to adapt to the chaotic nature of meteorological systems, thus significantly limiting the practical value of forecast results. To achieve probabilistic forecasting, early NWP ensemble models perturb initial states and model parameters, which entails extremely high computational costs and results in limited spatial resolution (Hu et al., 2023; Chen et al., 2023; Bi et al., 2023). Probabilistic machine learning methods integrate existing architectures with generative frameworks, including latent variable models (Hu et al., 2023; Oskarsson et al., 2024) and diffusion models (Tashiro et al., 2021; Price et al., 2023; Wen et al., 2023; Dong et al., 2025), aiming to directly learn generative models for producing ensemble members. However, the inherent characteristics of global weather station data deviate from the design assumptions of existing methods, creating a core challenge for direct migration to GSWF tasks: stations are geographically highly uneven (e.g., scarce in polar and desert regions) with an overall size far lower than satellite imagery, making it challenging for existing diffusion-based methods (typically jointly capture regular spatiotemporal patterns and random fluctuations) to model under limited, uneven data constraints (Yuan & Qiao, 2024).

To address the core challenges in GSWF: modeling the hierarchy and uncertainty of meteorological chaotic systems while adapting to the domain characteristics of global weather station data, we propose a novel Generative Hierarchical Mesh Network (GHMN). Specifically, to tackle insufficient hierarchical modeling, we first construct a multi-level graph on the spherical surface and establish a hierarchical system via bidirectional mapping between adjacent levels. Building on this, we map the latent state of irregularly distributed stations to the bottom-layer mesh nodes, then adopt a U-shaped information pathway: bottom-layer node states are upwardly encoded into top-layer nodes to accurately capture overall weather characteristics, which are then propagated downward to supplement local details. For uncertainty characterization, inspired by recent diffusion/flow matching loss functions (Li et al., 2024; Liu et al., 2025), we further introduce a flow matching framework to construct a lightweight generative prediction head that can generate multiple reasonable prediction results without presetting any prior distributions. Unlike existing schemes, we adopt the deterministic prediction results of the hierarchical mesh network (instead of latent space features) as the conditional input to the flow matching model. By synergizing a high-performance deterministic historical information modeling module for conditional mean prediction and core spatiotemporal pattern capture, with a lightweight uncertainty distribution modeling module that prioritizes post-mean uncertainty modeling over learning full data distributions from scratch, the design cuts learning complexity and is more suitable for global station data with uneven spatial coverage and limited size. To summarize:

- We propose a novel hierarchical mesh network for global station data, which effectively captures the hierarchical dependencies of meteorological systems.
- We design a lightweight flow matching head to replace the traditional deterministic modeling paradigm, generating multiple reasonable results without presetting any prior distributions.
- Experiments show that compared with spatiotemporal forecasting baselines, the proposed model achieves up to 11% performance improvement, with the advantage of probabilistic prediction.

2 Related Works

2.1 Deterministic Weather Forecasting

In recent years, machine learning-based meteorological models have emerged rapidly, forming a data driven paradigm without explicit physical parameterization. Representative works include: FourCastNet (Pathak et al., 2022), which employs Fourier neural operators for global field data;

GraphCast (Lam et al., 2023), which uses spherical graph neural networks for medium-range forecasting; Pangu-Weather (Bi et al., 2023), which introduces 3D convolutional architectures for volumetric prediction. However, as they operate on grid spaces, they may not be optimal for GSWF. In contrast, spatiotemporal series forecasting methods directly leverage historical weather observations and station spatial location information to predict future conditions, offering a more cost-effective alternative. Airphynet (Hettige et al., 2024) and Air-DualODE (Tian et al., 2025) capture spatiotemporal patterns among weather stations via graph convolution on a spatial graph. MIGN (Zheng et al., 2025), and CDPNet (Xu et al., 2025) first map the irregularly distributed stations onto a regular mesh, then perform message passing across mesh points to ensure the model learns patterns from both high- and low-density regions simultaneously. Yet these models usually only capture spatial correlation based on geographical proximity, while ignoring the global connectivity inherent to Earth's weather system. Moreover, most existing methods conduct single-scale forecasting, which hinders their ability to capture interactions among the multi-scale processes underpinning complex climate dynamics. Another mainstream solution is Transformer-based models. Representative works include: AirFormer (Liang et al., 2023), which introduces dartboard attention for spatial correlation modeling in air quality prediction; Corrformer (Wu et al., 2023), which organizes global stations into a hierarchical tree structure and performs sequential spatial correlation modeling for stations under each intermediate node; and S^2Transformer (Chen et al., 2025), which partitions the spatial graph into multiple subgraphs at different scales, leverages intra-subgraph attention to learn local spatial correlation, and aggregates nodes into subgraph representations to enable message passing across subgraphs at various scales. However, such models share a critical limitation: the efficacy of their core mechanisms hinges entirely on the quality of their spatial partitioning strategies. Regardless of whether clustering or graph partitioning algorithms are employed, these methods are prone to either unbalanced partitioning or information loss at partition boundaries, thereby compromising their capacity to capture spatial correlation. Additionally, STELLA (Fu et al., 2025) highlights the importance of spatial-temporal position embedding.

2.2 Probabilistic Weather Forecasting

As a chaotic system, weather inherently exhibits significant uncertainty, and accurate modeling of this uncertainty is crucial for enhancing prediction value. To sidestep this, models must accurately characterize the underlying distribution. Early machine learning-based numerical weather prediction ensemble methods either rely on ad-hoc initial condition perturbations (Chen et al., 2023; Bi et al., 2023) or fail to scale to practical spatial resolutions (Hu et al., 2023). While parametric density functions with predefined prior distributions (Wen et al., 2017; Woo et al., 2024) can be adopted to address uncertainty in spatiotemporal series forecasting, they tend to compromise the model's ability to learn complex underlying distributions, especially on global meteorological data characterized by high heterogeneity. Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020)—as advanced probabilistic generative models with flexible posterior probability approximation and robust generative capabilities—have been extended to spatiotemporal series forecasting (including the meteorological domain). Representative works include DiffSTG (Wen et al., 2023), which frames spatiotemporal graph prediction as a conditional diffusion generation task, integrating graph and temporal convolutions into denoising steps; Double-Diffusion (Dong et al., 2025), which embeds atmospheric physical principles as core constraints in both noise sampling and denoising; and DiffLiG (Li et al., 2025), which proposes a probabilistic diffusion selector to generate diverse correction candidates and select reliable outputs under uncertainty. To further reduce computational overhead, latent diffusion models operating in compressed latent spaces have been applied to weather forecasting (Gao et al., 2023; Zhuang & Duraisamy, 2025; Nguyen et al., 2025). However, global weather station data exhibits two inherent characteristics—limited sample size and uneven spatial distribution—that impede existing joint modeling paradigms from effectively learning historical conditional encodings and conducting reliable uncertainty modeling.

3 Preliminaries

Global Station Weather Probabilistic Forecasting. Global station weather data refers to multivariate time series data in which each series is associated with a station's spatial coordinates. Specifically, we define the spatial coordinates of all stations as $\boldsymbol{\lambda} \in \mathbb{R}^N$, $\boldsymbol{\phi} \in \mathbb{R}^N$ and the multivariate time series $\mathbf{X} \in \mathbb{R}^{N \times T \times C}$ records C-dimensional meteorological variables collected by N weather stations

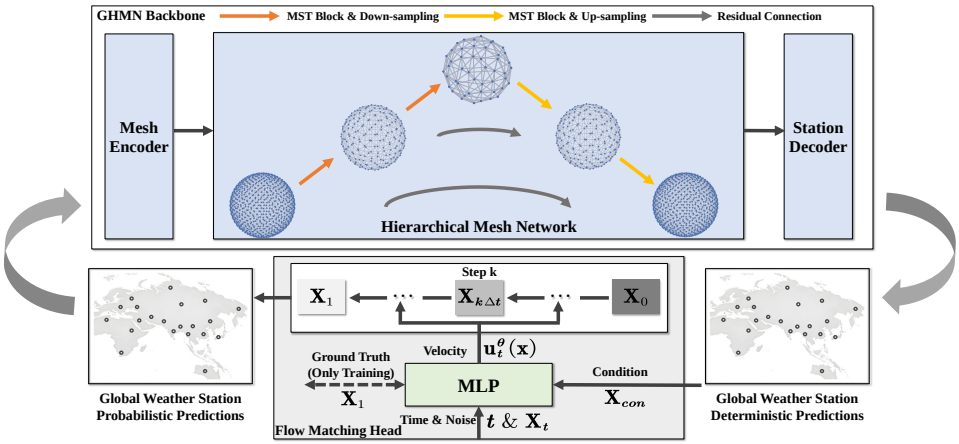


Figure 1: The framework of our proposed Generative Hierarchical Mesh Network. At the top, the hierarchical mesh network backbone captures core spatiotemporal patterns and generates deterministic predictions; at the bottom, the flow matching head models uncertainty and provides probabilistic predictions. Then, long-term predictions are derived via an autoregressive inference strategy.

over T time steps. Given the past T steps historical observations $\mathbf{X}_{t-T+1:t}$ along with the stations' longitude $\boldsymbol{\lambda}$ and latitude $\boldsymbol{\phi}$, the goal of global station weather probabilistic forecasting is to perform probabilistic prediction of meteorological variables for the future F time steps. This task can be generally formulated as a conditional probability distribution with historical observations and spatial coordinates as inputs:

$$p(\mathbf{X}_{t+1:t+F} \mid \mathbf{X}_{t-T+1:t}) = \mathcal{F}_\theta(\mathbf{X}_{t-T+1:t}, \boldsymbol{\lambda}, \boldsymbol{\phi}) \tag{1}$$

where $\mathcal{F}_\theta(\cdot)$ denotes the data-driven weather forecasting model parameterized by θ.

Flow Matching. The core goal of generative modeling is to learn the underlying generative probability distribution of data. Most current spatiotemporal generative forecasting tasks rely on diffusion models, while research on the flow-matching framework remains scarce. Derived from the concept of continuous normalizing flows, this framework achieves continuous transformation of samples from a source distribution $\mathbf{x}_0 \sim p_0$ to a target distribution $\mathbf{x}_1 \sim p_1$ by defining a time-dependent velocity field $\mathbf{u}_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, with its dynamic process satisfying the ordinary differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\psi}_t(\mathbf{x}) = \mathbf{u}_t\big(\boldsymbol{\psi}_t(\mathbf{x})\big),\ \boldsymbol{\psi}_0(\mathbf{x}) = \mathbf{x} \tag{2}$$

here, the velocity field $\mathbf{u}_t$ induces a flow $\boldsymbol{\psi}_t$. For any $t \in [0,1]$, $\boldsymbol{\psi}_t$ constructs a probability path p_t that interpolates between p_0 and p_1. Typically, we fit the target velocity field $\mathbf{u}_t$ by training a network $\mathbf{u}_t^\theta$ parameterized by θ, which is achieved by minimizing a loss function $\mathcal{L}(\theta)$, e.g., the mean squared error loss:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\mathbf{z}\sim p_1,\mathbf{x}\sim p_t(\cdot|\mathbf{z})} \left\|\mathbf{u}_t^\theta(\mathbf{x}) - \mathbf{u}_t(\mathbf{x})\right\|^2 \tag{3}$$

where $t \sim \text{Unif} = \text{U}[0,1]$. Lipman et al. (2022) proved the equivalence between this loss and the conditional loss:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{z}\sim p_1,\mathbf{x}\sim p_t(\cdot|\mathbf{z})} \left\|\mathbf{u}_t^\theta(\mathbf{x}) - \mathbf{u}_t(\mathbf{x}|\mathbf{z})\right\|^2 \tag{4}$$

By leveraging the conditional optimal-transport path and a source Gaussian, the objective can be formulated as:

$$\mathcal{L}_{\text{CFM}}^{\text{Gauss}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1} \left\|\mathbf{u}_t^\theta(\mathbf{x}_t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2 \tag{5}$$

where $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{x}_1 \sim p_1$, and $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$. Finally, we can generate new samples by applying a push-forward operation to the source Gaussian samples:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \mathbf{u}_t^\theta(\mathbf{x}_t)\Delta t,\ t \in [0,1] \tag{6}$$

4 METHODOLOGY

The framework of our proposed probabilistic forecasting model, dubbed Generative Hierarchical Mesh Network (GHMN), is shown in Figure 1. We first construct a hierarchical graph on the spherical surface (Section 4.1). We then map the latent state of irregular stations to the bottom-layer mesh nodes, encode it into the top-layer nodes to capture the global structure of weather states, and gradually supplement fine-grained features during downward propagation to yield deterministic predictions (at the top of Figure 1 and Section 4.2). For probabilistic forecasting, we introduce a lightweight flow matching-based generative head that takes the backbone's deterministic outputs as conditional inputs to generate multiple reasonable predictions (at the bottom of Figure 1 and Section 4.3).

4.1 HIERARCHICAL GRAPH

To address the uneven spatial distribution of global stations and the spatiotemporal continuity of atmospheric motion (He et al., 2020), we construct a hierarchical graph architecture based on the HEALPix mesh. Unlike traditional clustering or partitioning methods that are prone to unbalanced results and heavy hyperparameter tuning, mesh interpolation maps irregular station data to a regularized space. This enables the natural construction of a multi-scale structure through recursive subdivision, ensuring both physical consistency and computational efficiency. HEALPix (Gorski et al., 1999) is a hierarchical, equal-area, iso-latitude spherical pixelization method that partitions the sphere into uniform equal-area pixels, with a base resolution of twelve spherical quadrilateral pixels. Each pixel can be split into four child-grids via edge bisection; after k rounds of subdivision, a single original pixel is divided into 4^k pixels, leading to a total of 12×4^k global mesh nodes. The constructed hierarchical network consists of L levels graph $\mathcal{G} = \{G_i\}_{i=1}^{L}$, where $G_l = \{V_l, E_l\}$ represents the graph at the l-th level with $|V_l| = 12 \times 4^l$ and $E_l = \{(u, v) \mid u, v \in V_l,\ v \in \mathcal{N}_k(u)\}$, where $\mathcal{N}_k(u)$ denotes the set of the top k nearest neighbors of node u, and the distance is measured by the spherical spatial distance metric calculated from longitude and latitude. Additionally, bidirectional connections are established between nodes of adjacent levels based on the mesh generation relationship.

4.2 GHMN BACKBONE

Mesh Interpolate Encoder. To map irregular meteorological station data to regular mesh nodes, we adopt a graph neural network instead of traditional fixed-weight interpolation techniques (e.g., inverse distance weighting, kriging). First, following the setting of previous work (Wu et al., 2020; Chen et al., 2024), we reshape and transform the historical observation $\mathbf{X}_{t-T+1:t}$ into the input embeddings $\mathbf{X}_{\text{emb}} \in \mathbb{R}^{N\times T\times C\times D}$ with a linear map, where D is the embedding dimension. Next, we construct edges from station nodes ($|V_s| = N$) to bottom-layer mesh nodes (V_L) if the spatial distance between the nodes, computed from their respective longitude λ and latitude ϕ, is less than a predefined threshold δ. Based on this, we formalize the message passing mechanism as follows, where $\mathbf{A}_{\text{sm}} \in \mathbb{R}^{|V_L|\times|V_s|}$ denotes the station-mesh adjacency matrix and $\mathbf{W} \in \mathbb{R}^{D\times D}$ represent the learnable parameters.

$$\mathbf{H}_L = \mathbf{A}_{\text{sm}}\mathbf{X}_{\text{emb}}\mathbf{W} \tag{7}$$

Hierarchical Mesh Network. After constructing the hierarchical graph and mapping data to the bottom-layer regular mesh, we design a hierarchical mesh network to capture the hierarchical characteristics of meteorological systems. This structure shares core design principles with the widely used U-Net architecture (Ronneberger et al., 2015) in computer vision, while being extended to handle multivariate spatiotemporal data. Specifically, at the intra-level modeling stage, we design a Multivariate SpatioTemporal Block (MST Block) and deploy it separately during the upward and downward propagation phases within each hierarchical level. Given the input hidden state at level l, we first leverage self-attention and feed forward network to model dynamic variable correlations, then adopt GNN to capture spatial dependencies between nodes, and finally use a linear layer to model temporal dynamics, yielding the updated hidden state $\mathbf{H}_l$ for the current level.

For inter-level information propagation, we leverage the mesh-generated relationship to implement bidirectional propagation via down-sampling and up-sampling. Upward propagation starts from the bottom-layer mesh nodes, where information is aggregated to the top nodes via down-sampling for

capturing the global structure of weather states, with the formula as follows:

$$\mathbf{H}_{l-1}^{v} = \text{Conv1d}\left(\frac{1}{|\mathcal{C}(v)|}\sum_{u\in\mathcal{C}(v)}\mathbf{H}_{l}^{u}\right), \forall v \in V_{l-1} \tag{8}$$

where $\mathcal{C}(v)$ represents the child node set of v at level l, and Conv1d denotes the convolutional feature refinement operation that doubles the feature dimension. The upward propagation proceeds iteratively until it reaches the predefined top-level mesh nodes. Subsequently, downward propagation is performed: high-level abstract representations from the top-level nodes are propagated back to the bottom-level mesh nodes via up-sampling. This bidirectional propagation mechanism effectively integrates global meteorological trends and local fine-grained details.

$$\mathbf{H}_{l}^{u} \mathrel{+}= \text{Conv1d}\left(w_{v\to u}\mathbf{H}_{l-1}^{v}\right), \forall u \in V_l, v = \mathcal{P}(u) \tag{9}$$

where l denotes the current level, $\mathcal{P}(u)$ represents the unique parent node of u at level $l-1$, $w_{v\to u}$ is the distance-based interpolation weight, and Conv1d denotes the convolutional feature refinement operation that halves the feature dimension.

This backbone offers two core advantages: 1) Networks at different levels are equipped with independent parameters, distinguishing from multi-scale graph methods, where information processing across all spatial scales is confined to a single GNN layer (Lam et al., 2023). This enables the model to learn different representation update strategies for nodes across varying spatial scales; 2) It achieves natural spatially-aware dimensionality reduction, while fine-grained local details are progressively supplemented during downward propagation, balancing the modeling demands of global macro trends and regional micro features.

Station Interpolate Decoder. After hierarchical modeling, we map the bottom-layer mesh nodes' hidden states $\mathbf{H}_L$ back to the observation stations—serving as outputs for deterministic forecasting and conditional inputs for the flow matching head. This process follows a reverse mechanism of the encoder as follows, where $\mathbf{W} \in \mathbb{R}^{D\times 1}$ represent the learnable parameters.

$$\mathbf{X}_{\text{con}} = \mathbf{A}_{\text{sm}}^{T}\mathbf{H}_{L}\mathbf{W} \tag{10}$$

4.3 Flow Matching Head

After deriving deterministic predictions, we proceed to probabilistic modeling. Inspired by advances in diffusion and flow-matching loss functions (Li et al., 2024), we propose a flow-matching framework for probabilistic forecasting: it automatically fits target distributions without predefined parameters, generates raw prediction samples, and yields probabilistic results via sample statistics. Instead of latent space features, we take the deterministic predictions of the hierarchical mesh network as the conditional input to the flow-matching model, which offers two core advantages: 1) Unlike existing generative methods that simultaneously model spatiotemporal patterns and random fluctuations, we decouple forecasting into deterministic mean prediction and residual uncertainty estimation—this reduces learning complexity and better adapts to global station data with uneven spatial coverage and limited scale; 2) Focusing solely on uncertainty estimation, a lightweight network replaces complex denoising diffusion networks, effectively cutting inference costs.

Specifically, we decouple the core module $\mathcal{F}_\theta$ of the probabilistic forecasting model in Eq. 1 into two collaborative components. The first is the hierarchical mesh network $\mathcal{H}_\theta$ described earlier, which is optimized by minimizing the mean absolute error as follows:

$$\mathcal{L}_1 = \frac{\|\mathbf{X}_{\text{con}} - \mathbf{X}_1\|}{N \times F \times C} \tag{11}$$

where $\mathbf{X}_{\text{con}}$ is the deterministic prediction value in Eq. 10 and $\mathbf{X}_1$ is the groundtruth value ($\mathbf{X}_{t+1:t+F}$ in Eq. 1). The second component is $\mathcal{G}_\theta$—a lightweight trainable multi-layer perceptron (MLP) that serves as the velocity field network of the flow-matching framework, which is optimized as follows:

$$\mathcal{L}_2 = \mathbb{E}_{t,\,\mathbf{X}_0,\,\mathbf{X}_1}\left\|\mathcal{G}_\theta(\mathbf{X}_t,\, t,\, \mathbf{X}_{\text{con}}) - (\mathbf{X}_1 - \mathbf{X}_0)\right\|^2 \tag{12}$$

where $\mathbf{X}_0$ is a Gaussian noise, t is sampled from $\mathrm{U}[0,1]$, and $\mathbf{X}_t = t\mathbf{X}_1 + (1-t)\mathbf{X}_0$. During training, noisy samples $\mathbf{X}_t$ are sampled and jointly fed into the network together with time steps t;

Table 1: Global station weather forecasting performance comparison with deterministic spatiotemporal forecasting models. The best results are highlighted in **bold**, while the second-best results are underlined.

Variable	MAX TEMP		MIN TEMP		DEWP		SLP		WDSP		MXSPD	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
STGCN	9.74	2.22	9.44	2.11	9.25	2.11	24.15	3.42	8.60	2.01	20.63	3.27
MTGNN	8.94	2.17	8.10	2.03	7.80	1.99	22.06	3.28	8.14	1.96	19.52	3.18
HD-TTS	10.20	2.33	9.65	2.17	9.77	2.21	24.27	3.44	9.11	2.11	20.25	3.23
ReDyNet	10.33	2.26	10.85	2.30	10.81	2.32	24.15	3.40	8.75	2.06	20.95	3.28
GPS	10.91	2.42	10.37	2.27	11.13	2.50	25.24	3.57	8.79	2.04	20.89	3.30
iTransformer	9.75	2.25	9.33	2.11	9.24	2.15	24.72	3.47	8.82	2.06	21.03	3.32
PatchSTG	8.72	2.18	9.50	2.09	9.66	2.19	20.29	3.16	8.93	2.08	21.12	3.32
Corrformer	9.85	2.20	9.39	2.13	9.60	2.26	24.62	3.46	8.94	2.08	21.38	3.35
MIGN	8.47	2.09	8.01	1.99	7.92	1.97	20.09	3.12	8.38	1.98	19.73	3.19
STELLA	9.65	2.18	9.15	2.10	9.57	2.20	24.29	3.45	8.89	2.07	21.47	3.34
GHMN	**8.02**	**2.05**	**7.14**	**1.91**	**6.94**	**1.86**	**20.06**	**3.11**	**7.84**	**1.92**	**18.98**	**3.12**

the conditional prediction $\mathbf{X}_{con}$ is incorporated into the flow-matching network via the Adaptive Layer Normalization (AdaLN) mechanism (Peebles & Xie, 2023). Eventually, $\mathcal{H}_\theta$ and $\mathcal{G}_\theta$ are jointly optimized in an end-to-end manner with the total loss defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \mathcal{L}_2$. During the inference phase, we start from multiple distinct initial Gaussian noises and, based on Eq. 6, we set the step size as $\Delta t = 1/K$ (K is a hyperparameter), iteratively generate the velocity field to advance trajectory updates, and ultimately obtain a plausible set of prediction samples to achieve meteorological probabilistic forecasting.

5 Experiments

In this section, we evaluate our approach on global station weather datasets in both deterministic and probabilistic forecasting scenarios (Section 5.2 and Appendix B.1). Section 5.3 presents the ablation study, and Section 5.4, along with the Appendix B.2, describes the hyperparameter sensitivity analysis. To gain a deeper understanding of our model, we also conduct visualization experiments, which are provided in Appendix B.3.

5.1 Experimental Setup

Datasets. We evaluate our approach on the NOAA Global Surface Summary of the Day (GSOD) dataset. Following the experimental setup proposed by Zheng et al. (2025), we adopt 6 commonly observed daily meteorological variables, namely maximum temperature (MAX TEMP), minimum temperature (MIN TEMP), mean dew point (DEWP), mean sea level pressure (SLP), mean wind speed (WDSP), and maximum sustained wind speed (MXSPD). The dataset is publicly available via the official GSOD metadata portal and can be downloaded from the archive repository. Statistics of the classic global station weather forecasting dataset are presented in Table 4. Compared with the hourly dataset WEATHER-5K (Han et al., 2024), the daily-scale observations offer broader spatial coverage, with an average of over 9,000 stations versus approximately 6,000 stations for the hourly counterpart. The advantage in spatial coverage facilitates the learning of global-scale climate dynamics. The total count of stations corresponding to each variable is presented in Table 5.

Baselines. We compare our method with the following baselines: (1) GNN-based models: STGCN (Yu et al., 2018), MTGNN (Wu et al., 2020), HD-TTS (Marisca et al., 2024), ReDyNet (Gao et al., 2025); (2) Transformer-based models: GPS (Rampásek et al., 2022), iTransformer (Liu et al., 2024), PatchSTG (Fang et al., 2025); (3) Diffusion models: CSDI (Tashiro et al., 2021), DiffSTG (Wen et al., 2023); (4) models specifically designed for GSWF: Corrformer (Wu et al., 2023), MIGN (Zheng et al., 2025), STELLA (Fu et al., 2025). More details are provided in Appendix A.1.

Evaluation Metrics. We conduct a comprehensive comparison from the performance perspective, employing various evaluation metrics including mean absolute error (MAE), mean square error (MSE), symmetric extremal dependence index (SEDI), and continuous ranked probability score (CRPS). More details are provided in Appendix A.2.

Table 2: Global station weather forecasting performance comparison with probabilistic spatiotemporal forecasting models. The best results are highlighted in **bold**, while the second-best results are underlined.

Variable	MAX TEMP		MIN TEMP		DEWP		SLP		WDSP		MXSPD	
Metric	MAE	CRPS	MAE	CRPS	MAE	CRPS	MAE	CRPS	MAE	CRPS	MAE	CRPS
CSDI	5.43	2.60	8.19	2.62	6.55	2.30	3.48	2.62	3.74	1.76	4.01	2.11
DiffSTG	5.80	2.78	7.65	2.46	6.90	2.41	3.56	1.97	2.12	1.12	3.41	1.80
GHMN	**2.05**	**0.83**	**1.91**	**0.75**	**1.86**	**0.78**	**3.11**	**1.31**	**1.92**	**0.82**	**3.12**	**1.35**

Table 3: Ablation study. The best results are highlighted in **bold**.

Variable	MAX TEMP		MIN TEMP		DEWP		SLP		WDSP		MXSPD	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
w. IDW	9.57	2.32	9.41	2.26	9.43	2.25	24.03	3.47	8.84	2.03	20.79	3.25
w. single layer(l=3)	8.69	2.12	8.38	2.04	8.07	1.99	22.09	3.25	8.55	2.00	20.22	3.24
w. single layer(l=5)	8.53	2.11	8.06	2.03	7.81	1.97	22.07	3.24	8.45	2.02	20.19	3.25
w. MSE Head	8.28	2.08	7.56	1.95	7.61	1.94	21.31	3.20	8.06	1.96	19.45	3.17
w. Diffusion Head	8.70	2.12	7.83	1.99	7.58	1.93	21.66	3.23	8.43	1.98	19.92	3.18
GHMN	**8.02**	**2.05**	**7.14**	**1.91**	**6.94**	**1.86**	**20.06**	**3.11**	**7.84**	**1.92**	**18.98**	**3.12**

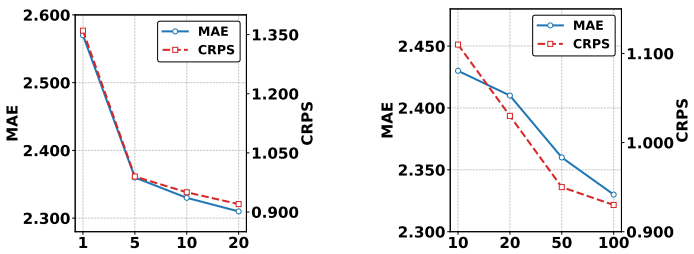
5.2 Performance Comparison

Point Forecasting. As summarized in Table 1, our GHMN consistently outperforms all baselines across all variables, benefiting from the accurate modeling of intrinsic meteorological system characteristics. In particular, GHMN achieves relative MSE reductions of 5.3%, 10.9%, and 11.0% for MAX TEMP, MIN TEMP, and DEWP, respectively. We further evaluate model performance on extreme weather regimes, where extreme events are defined as meteorological values exceeding the 90th upper quantile (e.g., heatwaves) or falling below the 10th lower quantile (e.g., winter storms). We adopt the SEDI metric (0–100%), a hit rate measure, to quantify the accuracy of extreme weather event identification, with higher SEDI scores indicating superior extreme weather prediction performance. As shown in Figure 3, GHMN maintains consistent superiority over all baselines across all variables, highlighting its robust generalization capability for extreme weather regimes. Additional results and analysis of autoregressive forecasting are provided in Appendix B.1. Although GHMN yields comparable point forecasting performance to baselines for SLP, it achieves notable improvements of 7.8% and 5.6% in extreme weather event identification and autoregressive forecasting results, respectively.

Probabilistic Forecasting. Beyond point forecasting, GHMN enables unique generative probabilistic predictions. As shown in Table 2, our GHMN substantially outperforms all diffusion baselines across all variables. Notably, these diffusion baselines even underperform deterministic baselines, as their coupled learning of spatiotemporal correlations and predictive uncertainty introduces excessive complexity. With limited training data, diffusion models fail to efficiently learn the true spatiotemporal data distribution. To tackle these issues, our GHMN decouples the deterministic spatiotemporal modeling module from the uncertainty estimation head, yielding an accurate, reliable, and high-performance probabilistic forecasting solution.

5.3 Ablation Study

We conduct an ablation study to verify the efficacy of the proposed modules, with the following model variants considered: (1) **w. IDW**: Inverse distance weighting (IDW) is used instead of a graph neural network to map irregular meteorological station data to regular mesh nodes. (2) **w. single layer (l=3)**: Hierarchical modeling is removed; multivariate spatiotemporal correlation is modeled only at the coarse scale (l=3, $V_l = 12 \times 4^3$). (3) **w. single layer (l=5)**: Hierarchical modeling is removed; multivariate spatiotemporal correlation is modeled only at the fine scale (l=5, $V_l = 12 \times 4^5$). (4) **w. MSE Head**: The flow matching head loss $\mathcal{L}_2$ is substituted with the MSE loss. (5) **w. Diffusion Head**: Replace the flow matching framework with a diffusion one, and the head loss $\mathcal{L}_2$ with a denoising diffusion-based objective. As shown in Table 3, the key findings are summarized as follows.



(a) Sampled Predictions (b) Sampling Steps

Figure 2: Parameter sensitivity analysis.

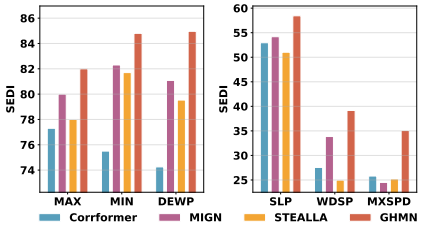


Figure 3: 90th percentile SEDI (%) of models specifically designed for global station weather forecasting. Note that the percentile here includes the lower bounds and upper bounds.

First, the sharp performance drop of **w. IDW** verifies that learnable weight mapping outperforms fixed distance-based IDW weighting, highlighting the necessity of adaptive spatial weighting. Second, the inferior performance of **w. single layer (l=3)** and **w. single layer (l=5)** demonstrates that a single scale—whether fine or coarse—is insufficient to capture the complex dynamic characteristics of meteorological systems. Finally, **w. MSE Head** exhibits poor performance and only yields a single deterministic prediction. In contrast, generative modeling supports downstream tasks by generating multiple plausible predictions that capture diverse future scenarios and improve decision reliability. While **w. Diffusion Head** enables prediction sampling; its performance with the diffusion-based objective is markedly lower than that of the flow matching loss.

5.4 Parameter Sensitivity Analysis

To validate the framework's robustness, we conduct a sensitivity analysis of two core hyperparameters: (1) the number of samples for statistical aggregation and (2) the flow-matching sampling step count, with results in Figure 2. As observed, increasing the sample count improves prediction accuracy—this suppresses random noise in individual predictions and better approximates the true data distribution. Meanwhile, fine-grained sampling steps enhance prediction precision by enabling refined trajectory optimization in the flow-matching space. Considering the trade-off between inference efficiency and performance, we adopt a practical configuration: 10 independent prediction samples, each generated via 50 flow-matching steps. Additionally, sensitivity analysis of adjacency matrix neighbor count and model embedding dimension is provided in Appendix B.2 for completeness.

6 Conclusions

In this paper, we propose a novel Generative Hierarchical Mesh Network for Global Station Weather Forecasting. Specifically, we construct a hierarchical spherical graph with bidirectional cross-level mapping, utilizing a U-shaped pathway to integrate local details and global meteorological patterns for accurate, deterministic predictions. The lightweight flow matching-based generative head takes these deterministic outputs as conditional inputs to generate plausible ensemble predictions without preset priors. This design enables the model to capture multi-scale hierarchical dependencies and the inherent uncertainty of chaotic meteorological systems, while adapting well to the uneven spatial distribution and limited size of global weather station data. Performance comparisons validate its superiority in both deterministic and probabilistic forecasting scenarios. Ablation studies confirm the effectiveness of key model components. For future work, we will extend the temporal span of the dataset and explore its potential to inspire paradigms for global meteorological foundation models.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for two purposes only: editing the manuscript to improve readability, i.e. polishing the grammar and wording of text drafted by the authors, and editing software code, i.e. auto-completion and refactoring suggestions while the authors implemented the method and the evaluation pipeline. We did not use generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or to assist in proving them, to propose or refine hypotheses, to design our methodology or experiments or to give feedback on them, to clean or reformat the dataset, to support qualitative or thematic data analysis, or to interpret the experimental results; the remaining tasks for which ICLR requires disclosure are not applicable to this work. We also did not use generative AI tools to retrieve, summarize, or select the related literature. All research ideas, the model design, the experimental protocol, and the reported results are the authors' own work. We have reviewed all AI-assisted work: all AI-assisted text was checked and rewritten by the authors, and all AI-assisted code was manually verified and tested for correctness by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure that our results are reproducible. The dataset is built entirely from the publicly available NOAA GSOD archive, whose access links, variable set, station counts, and train/validation/test splits by year are documented in Section 5.1 and Appendix A.3 (including Tables 4 and 5). The hierarchical HEALPix graph construction, the MST Block, and the flow matching head are fully specified in Section 4.1–4.3, and all default hyperparameters (embedding dimension, number of graph levels, neighbor count, number of samples and inference steps), the optimization setup, and the hardware used are listed in Appendix A.3. Baseline configurations follow the recommended settings in their official repositories, as described in Appendix A.1. All reported numbers are averaged over three independent runs. An anonymized implementation is available at `https://anonymous.4open.science/r/GHMN-334C`.

REFERENCES

Thomas Bailie, S Karthik Mukkavilli, Varvara Vetrova, and Yun Sing Koh. Hierarchical graph networks for accurate weather forecasting via lightweight training. *arXiv preprint arXiv:2510.22094*, 2025.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 2023.

Hongyi Chen, Xiucheng Li, Xinyang Chen, Yun Cheng, Jing Li, Kehai Chen, and Liqiang Nie. S^2transformer: Scalable structured transformers for global station weather forecasting. *arXiv preprint arXiv:2509.19648*, 2025.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. Fuxi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 2023.

Xiaodan Chen, Xiucheng Li, Xinyang Chen, and Zhijun Li. Structured matrix basis for multivariate time series forecasting with interpretable dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Vasudev Dehalwar, Akhtar Kalam, Mohan Lal Kolhe, and Aladin Zayegh. Electricity load forecasting for urban area using weather forecast information. In *IEEE International Conference on Power and Renewable Energy (ICPRE)*, 2016.

Hanlin Dong, Arian Prabowo, Hao Xue, and Flora D Salim. Double-diffusion: Diffusion conditioned diffusion probabilistic model for air quality prediction. *arXiv preprint arXiv:2506.23053*, 2025.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2025.

Yisong Fu, Fei Wang, Zezhi Shao, Boyu Diao, Lin Wu, Zhulin An, Chengqing Yu, Yujie Li, and Yongjun Xu. On the integration of spatial-temporal knowledge: A lightweight approach to atmospheric time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Qiang Gao, Zizheng Wang, Li Huang, Goce Trajcevski, Guisong Liu, and Xueqin Chen. Responsive dynamic graph disentanglement for metro flow forecasting. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Krzysztof M Gorski, Benjamin D Wandelt, Frode K Hansen, Eric Hivon, and Anthony J Banday. The healpix primer. *arXiv preprint astro-ph/9905275*, 1999.

Ismail Gultepe, R Sharman, Paul D Williams, Binbin Zhou, G Ellrod, P Minnis, S Trier, S Griffin, Seong S Yum, B Gharabaghi, et al. A review of high impact weather for aviation meteorology. *Pure and applied geophysics*, 2019.

Tao Han, Song Guo, Zhenghao Chen, Wanghan Xu, and Lei Bai. How far are today's time-series models from real-world weather forecasting applications? *arXiv preprint arXiv:2406.14399*, 2024.

Zhanjun He, Min Deng, Jiannan Cai, Zhong Xie, Qingfeng Guan, and Chao Yang. Mining spatiotemporal association patterns from complex geographic phenomena. *International Journal of Geographical Information Science*, 2020.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 2020.

Kethmi Hirushini Hettige, Jiahao Ji, Shili Xiang, Cheng Long, Gao Cong, and Jingyuan Wang. Airphynet: Harnessing physics-guided neural networks for air quality prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yuan Hu, Lei Chen, Zhibin Wang, and Hao Li. Swinvrnn: A data-driven ensemble forecasting model via learned distribution perturbation. *Journal of Advances in Modeling Earth Systems*, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Yuxiang Li, Yang Zhang, Guowen Li, Mengxuan Chen, Meng Jin, Fang Wang, Haohuan Fu, and Juepeng Zheng. Difflig: Diffusion-enhanced liquid graph with attention propagation for grid-to-station precipitation correction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning (ICML)*, 2025.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *International Conference on Machine Learning (ICML)*, 2024.

Michael J McPhaden, Stephen E Zebiak, and Michael H Glantz. Enso as an integrating concept in earth science. *Science*, 2006.

Tung Nguyen, Tuan Pham, Troy Arcomano, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Omnicast: A masked latent diffusion model for weather forecasting across time scales. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Qingjian Ni, Yuhui Wang, and Yifei Fang. Ge-stdgn: a novel spatio-temporal weather prediction model based on graph evolution. *Applied Intelligence*, 2022.

Joel Oskarsson, Tomas Landelius, Marc Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

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. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

Ladislav Rampásek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Yongjian Sun, Kefeng Deng, Kaijun Ren, Jia Liu, Chongjiu Deng, and Yongjun Jin. Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2024.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jindong Tian, Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, and Bin Yang. Air quality prediction with physics-guided dual neural odes in open systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Kingsley Eghonghon Ukhurebor, Charles Oluwaseun Adetunji, Olaniyan T Olugbemi, W Nwankwo, Akinola Samson Olayinka, C Umezuruike, and Daniel Ingo Hefft. Precision agriculture: Weather forecasting for future farming. In *Ai, edge and iot-based smart agriculture*. Elsevier, 2022.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, 2023.

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning (ICML)*, 2024.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 2023.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2020.

Chujie Xu, Yuqing Ma, Haoyuan Deng, Yajun Gao, Yudie Wang, Kai Lv, and Xianglong Liu. Continuous diffusive prediction network for multi-station weather prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Wanghan Xu, Kang Chen, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Extremecast: Boosting extreme value prediction for global weather forecast. *arXiv preprint arXiv:2402.01295*, 2024a.

Zhewen Xu, Xiaohui Wei, Jieyun Hao, Junze Han, Hongliang Li, Changzheng Liu, Zijian Li, Dongyuan Tian, and Nong Zhang. Dgformer: a physics-guided station level weather forecasting model with dynamic spatial-temporal graph neural network. *GeoInformatica*, 2024b.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Chengqing Yu, Guangxi Yan, Chengming Yu, Xinwei Liu, and Xiwei Mi. Mriformer: A multi-resolution interactive transformer for wind speed multi-step prediction. *Information Sciences*, 2024.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Zinan Zheng, Yang Liu, and Jia Li. Mesh interpolation graph network for dynamic and spatially irregular global weather forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yilin Zhuang and Karthik Duraisamy. Ladcast: A latent diffusion model for medium-range ensemble weather forecasting. *arXiv preprint arXiv:2506.09193*, 2025.

A Experiments Details

A.1 Baselines

We compare the proposed approach with the following advanced spatiotemporal forecasting baselines:

- STGCN: It utilizes graph convolutions to model local spatial correlation and 1D convolutions for temporal modeling in traffic forecasting.
- MTGNN: It builds an adaptive static global directed graph using learnable node embedding and aggregates information along spatial dimensions through mix-hop propagation.
- HD-TTS: A graph-based forecasting model that employs spatiotemporal downsampling to compute multi-scale representations, which are adaptively fused for prediction with missing data.
- ReDyNet: A dynamic graph model for metro flow forecasting that disentangles responsive dynamic graph structures via learning time-varying node embeddings and uses recurrent graph convolutions to capture evolving spatiotemporal dependencies.
- GPS: A scalable graph transformer model that models spatiotemporal dependencies via a generalized attention mechanism, enabling powerful and generalizable forecasting for spatiotemporal tasks.
- iTransformer: It embeds the whole time series into a spatial token and captures dynamic global spatial correlation using the self-attention mechanism.
- PatchSTG: It uses irregular spatial patching via KDTree and dual attention (depth and breadth) to capture local and global spatial correlations in large-scale spatiotemporal networks.
- CSDI: A conditional score-based diffusion model that leverages attention to capture temporal and feature dependencies, enabling probabilistic imputation, interpolation, and forecasting for time series data.
- DiffSTG: A non-autoregressive diffusion model that integrates STGNNs with a U-Net-based UGNet denoising network, capturing temporal patterns and spatial correlations to model uncertainties for spatiotemporal graph forecasting.
- Corrformer: It integrates multi-correlation mechanisms (spatial cross-correlation and temporal auto-correlation) in a learnable tree structure to model complex spatiotemporal dependencies for large-scale global station weather forecasting.
- MIGN: It maps irregular station data to a regular mesh via message passing and uses parametric spherical harmonics location embedding to boost spatial generalization for global weather forecasting.
- STELLA: A lightweight MLP model that fuses spatiotemporal knowledge through feature engineering and parameter-efficient design, delivering high accuracy and efficiency for large-scale atmospheric time series forecasting.

We obtain the code of baselines directly from their corresponding GitHub repositories. For the model- and training-related configurations, we follow the recommended settings provided in their code.

A.2 Evaluation Metrics

The performance evaluation metrics (MAE, MSE, SEDI, CRPS) are defined as follows:

$$\begin{aligned} \text{MAE} &= \frac{\sum_{n=1}^{N}\sum_{t=t_0}^{t_0+F-1}\sum_{c=0}^{C}|\hat{x}_{n,t,c}-x_{n,t,c}|}{N\times F\times C} \\ \text{MSE} &= \frac{\sum_{n=1}^{N}\sum_{t=t_0}^{t_0+F-1}\sum_{c=0}^{C}(\hat{x}_{n,t,c}-x_{n,t,c})^2}{N\times F\times C} \end{aligned} \tag{13}$$

where N is the number of stations, F is the forecast horizon, C is the number of variables; $x_{n,t,c}$ is the ground-truth value, and $\hat{x}_{n,t,c}$ is the corresponding prediction.

Table 4: Dataset statistics.

Dataset	Frequency	Time Span	Stations
WEATHER-5K	1 hour	2014-2023	5672
GSOD	1 day	1973-2025	over 9000

Table 5: Dataset variable statistic.

Variables Name	MAX TEMP	MIN TEMP	DEWP	WDSP	MXSPD	SLP
Stations	13260	13259	12709	13014	12895	9725

To quantify the model's capacity for extreme event identification, we adopt the symmetric extremal dependence index (SEDI) (Xu et al., 2024a). SEDI classifies predictions as extreme or normal based on ground-truth quantile thresholds and quantifies the overlap between predicted and ground-truth extreme events, defined as:

$$\text{SEDI}(p) = \frac{\sum \mathbb{I}(\hat{\mathbf{X}} < Q_{p,\text{lower}} \;\&\; \mathbf{X} < Q_{p,\text{lower}}) + \sum \mathbb{I}(\hat{\mathbf{X}} > Q_{p,\text{upper}} \;\&\; \mathbf{X} > Q_{p,\text{upper}})}{\sum \mathbb{I}(\mathbf{X} < Q_{p,\text{lower}}) + \sum \mathbb{I}(\mathbf{X} > Q_{p,\text{upper}})} \tag{14}$$

where $\mathbf{X}$ and $\hat{\mathbf{X}}$ denote the ground-truth and predicted values, respectively; $Q_{p,\text{lower}}$ and $Q_{p,\text{upper}}$ are the p-th lower and upper quantiles of the ground truth; and $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition holds, 0 otherwise). SEDI ranges in $[0, 100\%]$, with higher values indicating better performance in capturing extreme events.

For probabilistic forecasting, the continuous ranked probability score (CRPS) evaluates predictive distribution quality by measuring the discrepancy between predicted and empirical ground-truth cumulative distribution functions (CDFs), defined as:

$$\text{CRPS}(F, x) = \int_{-\infty}^{+\infty} \left[F(z) - \mathbb{I}(z \geq x)\right]^2 dz \tag{15}$$

where $F(z)$ is the predicted CDF and x is the observed ground-truth. It ranges in $[0, +\infty)$, with smaller values indicating better performance.

A.3 Implementation Details

We use data from 2017 to 2022 for model training, data from 2023 for validation, and data from 2024 for testing. Input data are normalized using the Z-score method during model training. To better reproduce our model, we summarize all the default hyperparameters as follows: Both the input embedding dimension and the hidden embedding dimension D are set to 64. The number of graph levels L is configured as 3 (with l ranging from 3 to 5). For graph construction, we employ a 10-neighbor approach. Specifically, the threshold ϵ and δ are set in accordance with the recommendations provided in the DCRNN framework (Li et al., 2017), and we further clip the neighborhood to the top 10 nearest neighbors to construct the adjacency matrices $\mathbf{A}_l$ and $\mathbf{A}_{sm}$. In addition, for flow matching inference, the number of samples S is set to 10, and the number of inference steps K is designated as 50. All experiments in this study are implemented using PyTorch Lightning and conducted on an NVIDIA RTX A6000 GPU with 48GB of memory. We run each experiment three times and report the average results.

B More Results

B.1 Auto-regressive Forecasting

Unlike existing time series forecasting methods that directly predict all future sequence values at once, mainstream research in machine learning weather prediction typically captures dynamic meteorological characteristics via short-time-step forecasting, then generates long-term future predictions through an autoregressive inference strategy. We thus conduct further autoregressive forecasting experiments

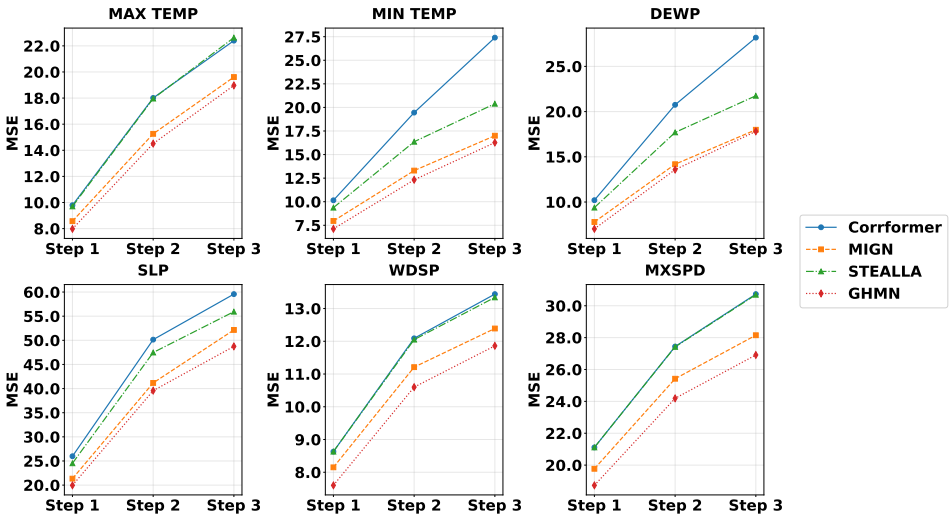
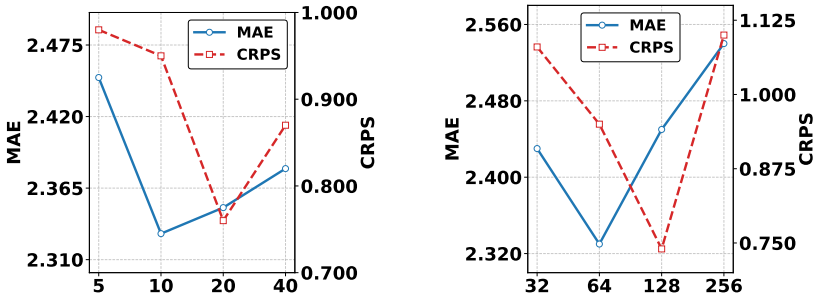


Figure 4: The autoregressive forecasting performance comparison of models specifically designed for global station weather forecasting.



(a) Neighbors (b) Embedding Dimension

Figure 5: More parameter sensitivity analysis.

to validate the performance of our model. As shown in Figure 4, GHMN consistently achieves state-of-the-art performance across all variables and all forecasting time steps. In particular, GHMN yields average relative MSE reductions of 5.6%, 5.3%, and 4.8% for SLP, WDSP, and MXSPD, respectively, which indicates that our model can effectively characterize the intrinsic properties of meteorological systems and mitigate the cumulative error in autoregressive forecasting.

B.2 More Parameter Sensitivity Analysis

We conduct a hyperparameter sensitivity analysis to examine two key factors: the neighbor count for constructing adjacency matrices $\mathbf{A}_l$ and $\mathbf{A}_{sm}$, and the embedding dimension D. As shown in Figure 5, the model achieves minimal loss with 10 or 20 neighbors. Using only 5 neighbors leads to significant performance degradation due to insufficient topological information in the adjacency matrices, while using 40 neighbors also impairs performance by introducing noise from distant or irrelevant nodes. For D, optimal performance is achieved at 64 or 128, with no further gains from increasing D. This confirms that a lightweight architecture is sufficient to capture the spatiotemporal patterns required for GSWF.

B.3 Visualization

Probabilistic Forecasting. As shown in Figure 6, our predictions well cover the observed values. By decoupling the spatiotemporal deterministic modeling module and the uncertainty estimation module, our point predictions closely align with the ground-truth mean values, while the probabilistic prediction intervals reliably cover the true values and capture extreme meteorological conditions.

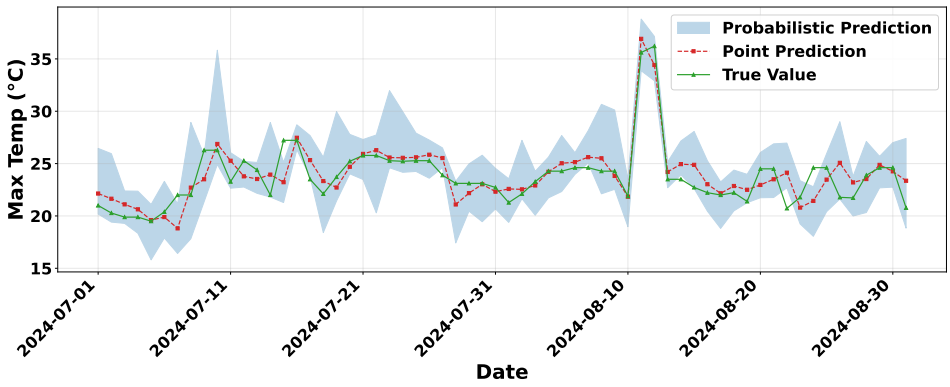


Figure 6: Visualization of probabilistic forecasting.

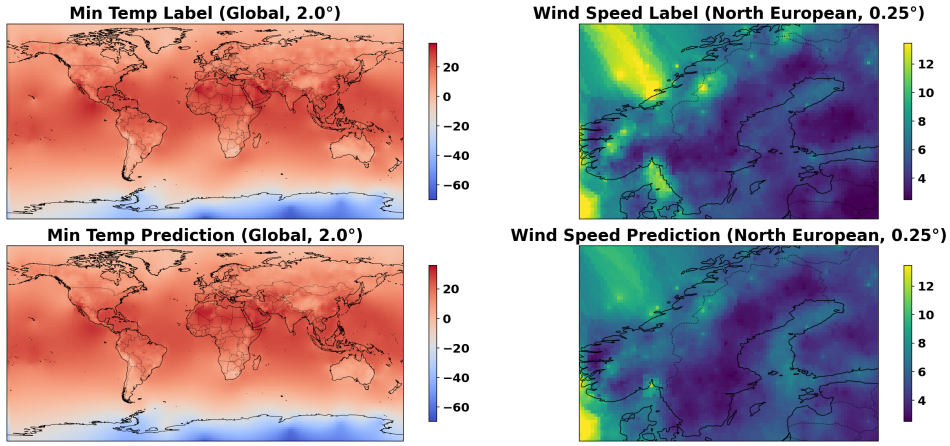


Figure 7: Visualization of global station weather forecasting.

Forecasting Results. As illustrated in Figure 7, our accurate modeling of the hierarchical characteristics of the global meteorological system enables our predictions to closely match the ground truth at both global and local scales. At the global scale, the predicted results clearly reflect the distribution patterns of distinct climatic temperature zones. At the local scale, particularly over Northern Europe, our model accurately depicts the northeast–southwest–oriented low wind speed zone—Norway's Scandinavian Mountains.

C Limitation

While GHMN offers a scalable framework for irregular station forecasting, we acknowledge certain limitations. Currently, capturing the absolute tails of chaotic distributions remains challenging due to high spatio-temporal heterogeneity and the inherent sparsity of extreme events. Furthermore, we plan to expand both the temporal coverage and spatial density of our datasets to provide a more robust and generalized paradigm for Global Weather Foundation Models.

D Broader Impact

In this paper, we propose a novel Generative Hierarchical Mesh Network to model the hierarchy and uncertainty of meteorological chaotic systems while adapting to the domain characteristics of global weather station data. Our research aims to make a positive contribution to the relevant community while ensuring no negative social impact.