Weather2K: A Multivariate Spatio-Temporal Benchmark Dataset for Meteorological Forecasting Based on Real-Time Observation Data from Ground Weather Stations

Xun Zhu[1, †] **Yutong Xiong**[1, †] **Ming Wu**[1, ‡] **Gaozhen Nie**[2] **Bin Zhang**[1] **Ziheng Yang**[1]
[1]Beijing University of Posts and Telecommunications, China
[2]National Meteorological Center, China Meteorological Administration, China

Abstract

Weather forecasting is one of the cornerstones of meteorological work. In this paper, we present a new benchmark dataset named Weather2K, which aims to make up for the deficiencies of existing weather forecasting datasets in terms of real-time, reliability, and diversity, as well as the key bottleneck of data quality. To be specific, our Weather2K is featured from the following aspects: 1) Reliable and real-time data. The data is hourly collected from 2,130 ground weather stations covering an area of 6 million square kilometers. 2) Multivariate meteorological variables. 20 meteorological factors and 3 constants for position information are provided with a length of 40,896 time steps. 3) Applicable to diverse tasks. We conduct a set of baseline tests on time series forecasting and spatio-temporal forecasting. To the best of our knowledge, our Weather2K is the first attempt to tackle weather forecasting task by taking full advantage of the strengths of observation data from ground weather stations. Based on Weather2K, we further propose Meteorological Factors based Multi-Graph Convolution Network (MFMGCN), which can effectively construct the intrinsic correlation among geographic locations based on meteorological factors. Sufficient experiments show that MFMGCN improves both the forecasting performance and temporal robustness. We hope our Weather2K can significantly motivate researchers to develop efficient and accurate algorithms to advance the task of weather forecasting. The dataset can be available at https://github.com/bycnfz/weather2k/.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. † **Equal contribution. ‡ Corresponding author: wuming@bupt.edu.cn**

1 INTRODUCTION

The changing weather has been profoundly affecting people's lives since the beginning of mankind. Weather conditions play a crucial role in production and economic industries such as transportation, tourism, agriculture and energy. Therefore, reliable and efficient weather forecasting is of great economic, scientific and social significance. The weather forecasting task deserves extensive attention.

Meteorological factors, such as temperature, humidity, visibility, and precipitation, can provide strong support and historical information for researchers to analyze the variation tendency of weather. For the past few decades, Numerical Weather Prediction (NWP) is the widely used traditional method, which utilizes physical models to simulate and predict meteorological dynamics in the atmosphere or on the Earth's surface (Müller and Scheichl, 2014). However, the prediction of NWP may not be accurate enough due to the uncertainty of the initial conditions of the differential equation (Tolstykh and Frolov, 2005), especially in complex atmospheric processes. In addition, NWP has high requirements on computing power.

In recent years, meteorological researchers have achieved considerable breakthroughs and successes in introducing data-driven approaches, most prominently deep learning methods, to the task of weather forecasting. Data-driven approaches exploit the historical meteorological observation data aggregated over years to model patterns to learn the input-output mapping. In the task of time series forecasting, Transformers have shown great modeling ability for long-term dependencies and interactions in sequential data benefiting from the self-attention mechanism. In the more challenging task of spatio-temporal forecasting, the classical Convolutional Neural Networks (CNNs) which work well on regular grid data in Euclidean domain have been greatly challenged to handle this problem due to the irregular sampling of most spatio-temporal data. The Graph Neural Networks (GNNs), which have already been extensively applied to traffic forecasting, yield effective and efficient performance by properly treating the variables and

the connections among them as graph nodes and edges, respectively. However, studies focusing on the field of meteorology are relatively scarce, while the demand for weather forecasting is increasing dramatically.

For data-centric deep learning method, the performance of the model heavily depends on the quality of the available training data. High-quality benchmark datasets can serve as "catalysts" for quantitative comparison between different algorithms and promote constructive competition (Ebert-Uphoff et al., 2017). In the field of meteorological science, reanalysis dataset cannot ensure the authenticity of the data. Remote sensing dataset cannot reliably reflect complexities of near-surface weather conditions. Fusion dataset cannot guarantee real-time performance. Inadequacies of existing datasets also include diversity of meteorological factors and applicable tasks. In this work, we present a new benchmark dataset named Weather2K, aiming at advancing the progress on weather forecasting tasks based on deep learning methods. As far as we know, our Weather2K is the first attempt to tackle the challenging weather forecasting task by entirely using the observation data from the ground weather stations. Concretely, our Weather2K has the following characteristics:

Reliable and Real-time Data The raw data is built on the one-hourly observation data of 2,130 ground weather stations from the China Meteorological Administration (CMA), which can be updated hourly in real-time. We have spent considerable time on data processing to ensure continuous temporal coverage and consistent data quality.

Multivariate Meteorological Variables Based on meteorological consideration, 20 important near-surface meteorological factors and 3 time-invariant constants for position information are provided in the Weather2K.

Applicable to Diverse Tasks We provide two versions of Weather2K for different application directions: time series forecasting and spatio-temporal forecasting. Well-arranged spatio-temporal sequences can be easily selected and accessed for different tasks of weather forecasting.

As a benchmark, we conduct a set of baseline tests on two application directions to validate the performance of our Weather2K. 4 representative transformer-based models and 8 state-of-the-art spatio-temporal GNN models are set up for comparison in time series forecasting and spatio-temporal forecasting, respectively. We further propose Meteorological Factors based Multi-Graph Convolution Network (MFMGCN). Considering utilizing multivariate meteorological information, MFMGCN fuses 4 static graphs representing different types of spatio-temporal information and 1 dynamic graph to model correlations among geographic locations, and also uses a complete convolutional structure followed time-space consistency. We validate both the improvement and temporal robustness of MFMGCN on Weather2K through extensive experiments.

We hope our Weather2K can significantly motivate more researchers to develop efficient and accurate algorithms and help ease future research in weather forecasting task.

2 PREVIOUS WORK

2.1 Meteorological Science Datasets

There are many types and sources of data commonly used in the meteorological science community. Tropical Rainfall Measuring Mission (TRMM) precipitation rate dataset (Kummerow et al., 2000) and the Integrated Multi-satellite Retrievals for GPM (IMERG) (Huffman et al., 2015) are developed on remote sensing data from satellites. Multi-Source Weighted-Ensemble Precipitation (MSWEP) (Beck et al., 2017) merges satellite and reanalysis data. The ERA5 (Hersbach et al., 2020) is the most advanced global reanalysis product created by the European Center for Medium Weather Forecasting (ECMWF). The Global Land Data Assimilation System (GLDAS) (Rodell et al., 2004) and the Canadian Land Data Assimilation System (CaLDAS) (Carrera et al., 2015) are well-known gridded datasets at global and regional scales, respectively. It is worth noting that the China Meteorological Forcing Dataset (CMFD) (He et al., 2020) is the first gridded near-surface meteorological dataset with high spatio-temporal resolution. Differently, our Weather2K is the first truly meaningful meteorological science dataset entirely based on the observation data from ground weather stations in China.

2.2 Weather Forecasting Datasets

In the task of weather forecasting, different datasets are constructed in various ways. And the number of involved meteorological factors varies widely. ExtremeWeather (Racah et al., 2017) is an image dataset for detection, localization, and understanding of extreme weather events. CloudCast (Nielsen et al., 2021) is a satellite-based image dataset for forecasting 10 different cloud types. Jena Climate (https://www.bgc-jena.mpg.de/wetter/) is made up of 14 meteorological factors recorded over several years at the weather station of the Max Planck Institute for Biogeochemistry. Climate Change (http://berkeleyearth.org/data/) provided by the Berkeley Earth focuses on global land and ocean temperature data. WeatherBench (Rasp et al., 2020) is a benchmark dataset for data-driven medium-range weather forecasting, specifically 3–5 days. Notably, our Weather2K is a spatio-temporal dataset with 20 optional multivariate meteorological factors, which can be applied to different directions of weather forecasting.

2.3 Transformers in Time Series Forecasting

Due to the special characteristics of time series forecasting task, the innovation of Transformer (Vaswani et al.,

Table 1: Definitions and physical descriptions of the variables in the Weather2K dataset.

Long Name	Short Name	Unit	Physical Description
Latitude	lat	(°)	The latitude of the ground observation station
Longitude	lon	(°)	The longitude of the ground observation station
Altitude	alt	(m)	The altitude of the air pressure sensor
Air pressure	ap	hpa	Instantaneous atmospheric pressure at 2 meters above the ground
Water vapor pressure	wvp	hpa	Instantaneous partial pressure of water vapor in the air
Air Temperature	t	(°C)	Instantaneous temperature of the air at 2.5 meters above the ground where sheltered from direct solar radiation
Maximum / Minimum temperature	mxt / mnt	(°C)	Maximum / Minimum air temperature in the last one hour
Dewpoint temperature	dt	(°C)	Instantaneous temperature at which the water vapor saturates and begins to condense
Land surface temperature	st	(°C)	Instantaneous temperature of bare soil at the ground surface
Relative humidity	rh	(%)	Instantaneous humidity relative to saturation at 2.5 meters above the ground
Wind speed	ws	(ms^{-1})	The average speed of the wind at 10 meters above the ground in a 10-minute period
Maximum wind speed	mws	(ms^{-1})	Maximum wind speed in the last one hour
Wind direction	wd	(°)	The direction of the wind speed. (Wind direction is 0 if wind speed is less than or equal to 0.2)
Maximum wind direction	mwd	(°)	Maximum wind speed's direction in the last one hour
Vertical visibility	vv	(m)	Instantaneous vertical visibility
Horizontal visibility in 1 min / 10 min	hv1 / hv2	(m)	1 / 10 minute(s) mean horizontal visibility at 2.8 meters above the ground
Precipitation in 1h / 3h / 6h / 12h / 24h	p1 / p2 / p3 / p4 / p5	(mm)	Cumulative precipitation in the last 1 / 3 / 6 / 12 / 24 hour(s)

2017) and its variants based on the self-attention mechanism shows great capability in sequential data. MetNet (Sønderby et al., 2020) uses axial self-attention to aggregate the global context from radar and satellite data for precipitation forecasting. LogTrans (Li et al., 2019b) and Reformer (Kitaev et al., 2020) propose the LogSparse attention and the local-sensitive hashing attention to reduce the complexity of both memory and time, respectively. AST (Wu et al., 2020a) utilizes a generative adversarial encoder-decoder pipeline. Autoformer (Wu et al., 2021) proposes a seasonal-trend decomposition architecture with an auto-correlation mechanism. Informer (Zhou et al., 2021) utilizes Kullback-Leibler divergence based the ProbSparse attention. Pyraformer (Liu et al., 2021) captures different ranges of temporal dependencies in a compact, multi-resolution fashion. FEDformer (Zhou et al., 2022) uses Fourier transform and wavelet transform to consider the characteristics of time series in the frequency domain.

2.4 Spatio-temporal Graph Neural Networks.

Spatio-temporal forecasting models are mostly based on GNNs due to their ability to learn representations of spatial irregular distributed signals by aggregating or diffusing messages from or to neighborhoods. GNNs have already brought a huge boost in traffic forecasting, such as STGCN (Yu et al., 2017), DCRNN (Li et al., 2017), MST-GCN (Guo et al., 2019), ASTGCN (Guo et al., 2019), TGCN (Zhao et al., 2019), AGCRN (Bai et al., 2020), and GMAN (Zheng et al., 2020). Moreover, LRGCN (Li et al., 2019a), 2s-AGCN (Shi et al., 2019) and MPNN (Panagopoulos et al., 2021) highlight the usefulness of GNNs in path failure in a telecommunication network, skeleton-based action recognition and epidemiological prediction, respectively. In the field of meteorological science, DeepSphere(Defferrard et al., 2020) introduces a method based on a graph representation of the sampled spherical meteorological data. PM2.5-GNN (Wang et al., 2020) and CLCRN (Lin et al., 2022) are used to predict PM2.5 concentrations and meteorological factors including temperature, cloud cover, humidity, and surface wind component. In general, researches on the application of GNNs to weather forecasting are relatively scarce.

3 THE DATASET OF WEATHER2K

The raw data comes from CMA's observation data of China's ground weather stations. Observation instruments and sensors of different weather stations all follow the standard of *Specifications for surface meteorological observation—General* (GB/T 35221-2017) and *Quality control of surface meteorological observation data* (QX/T 118-2010). While collecting the original data, basic quality control (QC) technologies are first used for different meteorological factors to correct for sensor noise, bias, and external factors. Focusing on missing values, defaults, outliers, we performed rigorous filtering and processing to ensure the quality of our dataset. The main steps of our data processing procedure are provided in Supplementary Material C.1.

3.1 Dataset Statistics

Weather2K contains the observation data from 2,130 ground weather stations throughout China, covering an area of more than 6 million square kilometers. The data are available from January 2017 to August 2021 with a temporal resolution of 1 hour, in which we can ensure the use of unified meteorological observation instruments and sensors for data collection. All stations use CST time (UTC + 8h). Based on the consideration of building a representative and comprehensive dataset for meteorological forecasting, Weather2K contains 3 time-invariant constants: latitude, longitude, and altitude to provide position information and 20 important meteorological factors with a length of 40,896 time steps. The definitions and physical descriptions of the chosen variables are listed in Table 1. The current

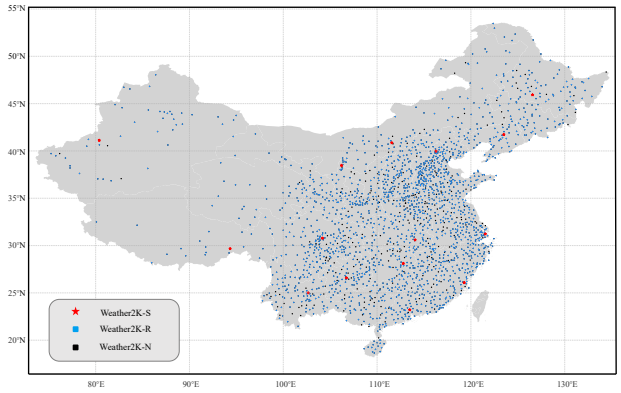

Figure 1: The distribution of the 2,130 ground weather stations that supply observation data for Weather2K.

version of our open source Weather2K is Weather2K-R, which contains 1866 ground weather stations and is stored in Numpy format. Weather2K-N, which contains the rest of the stations, is temporarily reserved for confidentiality reasons. In addition, we further provide a special version of the Weather2K (Weather2K-S), which contains 15 representative ground weather stations distributed in different regions and is stored in CSV format files. As shown in Figure 1, the red stars, blue squares, and black squares represent the stations that make up Weather2K-S, Weather2K-R, and Weather2K-N, respectively. See more detailed statistics on Weather2K in Supplementary Material C.2.

3.2 Comparison with Other Datasets

Our Weather2K is a novel attempt to solve the task of weather forecasting by entirely using the observation data from the ground weather stations. Although CMFD (He et al., 2020) also uses approximately 700 stations of CMA as its primary data source, it must incorporate several gridded datasets from remote sensing and reanalysis. Jena Climate (https://www.bgc-jena.mpg.de/wetter/) only records the data of 1 weather station in Jena, Germany.

Most of the existing multivariate datasets cover a long time range. For example, CMFD (He et al., 2020) and WeatherBench (Rasp et al., 2020) are both from 1979 to 2018. Limited by the purpose of data collection using unified meteorological observation instruments and sensors, our Weather2K begins in January 2017 and is ongoing.

The most attractive advantage of our Weather2K is the real-time and reliability of the data. The combination of high precision sensor and high-quality data control ensures the accuracy of the data to the greatest extent. The efficient and practical data processing can obtain hourly real-time data. In comparison, CMFD (He et al., 2020), ERA5 (Hersbach et al., 2020), and WeatherBench (Rasp et al., 2020) have a latency of at least 3 months because they are limited by the release time of the raw data.

In terms of meteorological factors, our Weather2K provides a total number of 20 optional variables. By contrast, Climate Change (http://berkeleyearth.org/data/) focuses on meteorological factors related to temperature. CMFD (He et al., 2020) provides 7 near-surface meteorological factors. WeatherBench (Rasp et al., 2020) contains 14 meteorological factors, 8 of which have multiple vertical levels. Moreover, our Weather2K can be applied to weather forecasting tasks in different directions. Weather2K-S and Weather2K-R are designed for the task of time series forecasting and spatio-temporal forecasting, respectively, while existing datasets mainly focus on solving a single task.

4 APPLICATIONS

In this section, we mainly introduce the application of the Weather2K in two important directions of weather forecasting, namely time series forecasting and spatio-temporal forecasting. For the convenience of expression, the temperature, visibility, and humidity mentioned in section 4 and section 5 refer to air temperature, horizontal visibility in 10 min, and relative humidity, respectively. These 3 meteorological factors are also selected as our forecasting targets. See more details about baseline models and implementation in Supplementary Material D.1 and D.2, respectively.

4.1 Time Series Forecasting

Weather2K-S is the special version of Weather2K designed for the task of time series forecasting, which contains 15 representative and geographically diverse ground weather stations distributed throughout China. We conduct three types of experiments on Weather2K-S: univariate forecasting, multivariate forecasting, and multivariate to univariate forecasting. 4 representative transformer-based baselines are set up for comparison, i.e., Transformer (Vaswani et al., 2017), Reformer (Kitaev et al., 2020), Informer (Zhou et al., 2021), and Autoformer (Wu et al., 2021). In meteorological science, short-term, medium-term, and long-term forecasting can be divided into 0-3 days, 3-15 days, and more than 15 days, respectively. For better comparison, we set the prediction time lengths for both training and evaluation to 24 (1 day), 72 (3 days), 168 (7 days), 336 (14 days), and 720 (30 days) steps, while the input time length is fixed to 72 (3 days) steps. We evaluate the performance of predictive models by two widely used metrics, including Mean Square Error (MSE) and Mean Absolute Error (MAE).

Table 2 shows the forecasting performance of temperature, visibility, and humidity under different baselines and prediction lengths. In univariate forecasting and multivariate forecasting, both the input and output are the single and multiple meteorological factors we specify, respectively. By comparison, we can intuitively find that multivariate forecasting is more complex than univariate forecasting. Meteorological forecasting becomes more challenging as

Table 2: Univariate results and multivariate results under different baselines with different prediction lengths of 24, 72, 168, 336, and 720 steps. The input time length is 72 steps. The reported results of mean and standard deviation are obtained through experiments under 15 different ground weather stations in Weather2K-S. Results with **bold** and underlines are the best and worst performance achieved by baselines, respectively.

Models		Transformer		Reformer		Informer		Autoformer	
Factors	**Metrics**	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Temperature	24	**0.0740**±0.0214	**0.2000**±0.0284	0.0897±0.0241	0.2306±0.0289	0.0758±0.0229	0.2031±0.0299	0.0978±0.0290	0.2367±0.0342
	72	**0.1246**±0.0396	**0.2656**±0.0440	0.1294±0.0393	0.2797±0.0400	0.1272±0.0355	0.2711±0.0386	0.1384±0.0474	0.2832±0.0464
	168	0.1560±0.0513	0.2999±0.0502	**0.1431**±0.0528	**0.2926**±0.0529	0.1699±0.0529	0.3162±0.0660	0.1574±0.0583	0.3002±0.0544
	336	0.1709±0.0635	0.3148±0.0587	**0.1451**±0.0562	**0.2947**±0.0576	0.2118±0.0660	0.3586±0.0508	0.1655±0.0630	0.3092±0.0575
	720	0.1603±0.0671	0.3058±0.0630	**0.1481**±0.0532	**0.2968**±0.0544	0.2119±0.0665	0.3637±0.0536	0.2011±0.0602	0.3453±0.0510
Visibility	24	0.7105±0.1155	0.6536±0.1151	**0.6623**±0.1108	**0.6288**±0.1175	0.7007±0.1198	0.6390±0.1218	0.8014±0.1589	0.6929±0.1013
	72	0.8722±0.1945	0.7475±0.1487	**0.8304**±0.1843	**0.7440**±0.1526	0.8847±0.2079	0.7503±0.1400	0.9350±0.2215	0.7585±0.1336
	168	0.9153±0.1969	0.7802±0.1439	**0.9083**±0.2380	**0.7795**±0.1441	0.9428±0.1942	0.7943±0.1523	0.9911±0.2273	0.7834±0.1289
	336	0.9952±0.2766	0.8059±0.1855	**0.9219**±0.2713	**0.7810**±0.1706	1.0232±0.2336	0.8388±0.1720	0.9919±0.2311	0.7832±0.1415
	720	0.9963±0.2570	0.8090±0.1755	**0.9506**±0.2853	**0.7946**±0.1763	1.0235±0.2618	0.8438±0.0789	1.0651±0.2591	0.8145±0.1553
Humidity	24	0.3746±0.0820	0.4607±0.0548	**0.3618**±0.0725	**0.4529**±0.0438	0.3749±0.0848	0.4563±0.0529	0.4291±0.1058	0.4941±0.0625
	72	0.5228±0.1310	0.5603±0.0782	**0.4932**±0.1074	**0.5455**±0.0621	0.5232±0.1234	0.5619±0.0769	0.5511±0.1209	0.5721±0.0651
	168	0.5935±0.1533	0.6004±0.0814	**0.5421**±0.1254	**0.5759**±0.0688	0.6221±0.1567	0.6163±0.0884	0.6111±0.1304	0.6042±0.0651
	336	0.6107±0.1643	0.6142±0.0815	**0.5761**±0.1456	**0.5938**±0.0749	0.6965±0.1845	0.6653±0.1051	0.6514±0.1567	0.6258±0.0737
	720	0.6177±0.1638	0.6146±0.0833	**0.5864**±0.1399	**0.6009**±0.0737	0.7979±0.1538	0.7285±0.0789	0.7145±0.1631	0.6599±0.0741
Multivariate	24	0.7134±0.1796	0.4597±0.0418	**0.6342**±0.1700	**0.4177**±0.0178	0.6838±0.1777	0.4381±0.0338	0.7525±0.1968	0.4868±0.0300
	72	0.8143±0.2071	0.5179±0.0469	**0.7616**±0.1877	**0.4914**±0.0330	0.8142±0.1967	0.5128±0.0391	0.8585±0.2132	0.5390±0.0314
	168	0.8462±0.2076	0.5261±0.0430	**0.8029**±0.1979	**0.5075**±0.0339	0.8641±0.2004	0.5543±0.0311	0.8945±0.2160	0.5455±0.0390
	336	0.8507±0.2044	0.5450±0.0396	**0.8251**±0.2057	**0.5173**±0.0337	0.9040±0.2129	0.5750±0.0307	0.9298±0.2222	0.5575±0.0381
	720	0.8967±0.2308	0.5543±0.0446	**0.8501**±0.2279	**0.5240**±0.0335	0.9242±0.2241	0.5833±0.0329	0.9379±0.2553	0.5731±0.0432

Table 3: Univariate results with 8 spatio-temporal GNN models in prediction time length of 12 steps on Weather2K-R. The input time length is 12 steps. Results with **bold** and underlines are the best and worst performance, respectively.

Factors	**Metrics**	STGCN	ASTGCN	MSTGCN	DCRNN	GCGRU	TGCN	A3TGCN	CLCRN
Temperature	MAE	3.1297	2.6942	3.6827	5.0457	1.8823	2.6429	2.9372	**1.6498**
	RMSE	4.7154	3.6383	4.9195	6.4070	2.7491	3.6189	4.0045	**2.4328**
Visibility	MAE	7.7568	5.7311	5.9053	7.6464	4.4863	6.0115	5.7731	**4.1252**
	RMSE	9.0896	7.5523	7.8016	9.0035	6.8236	8.4893	7.5974	**6.4660**
Humidity	MAE	9.8479	12.0332	14.4780	16.3467	**7.7008**	10.7045	12.1034	7.8056
	RMSE	14.5123	15.9847	18.7188	21.2517	**10.9422**	14.3170	15.9838	11.1326

the length of the predicted time series increases. On Weather2K-S, Reformer (Kitaev et al., 2020) have the best performance most of the time due to its efficient locality-sensitive hashing attention. The poor performance of Autoformer (Wu et al., 2021) may be that it attempts to construct a series-level connection based on the process similarity derived by series periodicity, whereas meteorological data in Weather2K-S change dynamically and irregularly.

We further explore to use the multivariate data to forecast the target univariate meteorological factor. However, multivariate information is underutilized in existing transformer-based time series forecasting baselines. The performance of multivariate to univariate forecasting is provided in Supplementary Material D.3. To make the evaluation more complete and informative, we also present the univariate forecasting results of the persistent model and some classical nonparametric methods in Supplementary Material D.4.

4.2 Spatio-Temporal Forecasting

Unlike time series forecasting that only targets a single selected ground weather station, spatio-temporal forecasting that can cover the entire region is more challenging and meaningful in the meteorological domain due to the highly nonlinear temporal dynamics and complex location-characterized patterns. We conduct experiments for univariate forecasting on Weather2K-R. We select 8 state-of-the-art spatio-temporal GNN models as baselines: STGCN (Yu et al., 2017), ASTGCN (Guo et al., 2019), MSTGCN (Guo et al., 2019), DCRNN (Li et al., 2017), GCGRU (Seo et al., 2018), TGCN (Zhao et al., 2019), A3TGCN (Bai et al., 2021), and CLCRN (Lin et al., 2022). We construct a neighbor graph based on distance as one of the GNN models' inputs by utilizing three time-invariant constants that provide position information including latitude, longitude, and altitude. We set both the input time length and prediction time length as 12 steps. We evaluate the performance by two commonly used metrics in spatio-temporal forecasting, including MAE, Root Mean Square Error (RMSE).

Table 3 summarizes the univariate forecasting results of 8 spatio-temporal GNN models based on Weather2K-R. Since most of the compared models are established for the task of traffic forecasting, it is not surprising that they show significant degradation in performance for meteorological

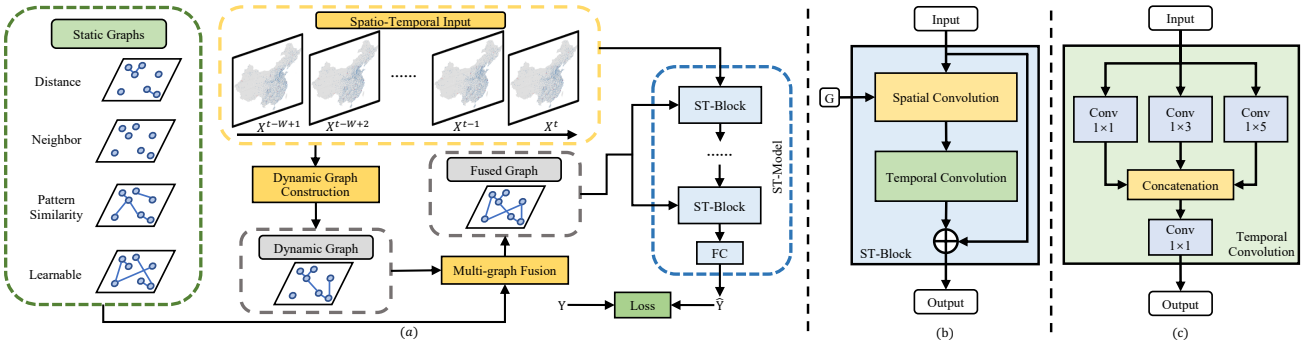

Figure 2: (a) The overview of Meteorological Factors based Multi-Graph Convolution Network (MFMGCN). (b) The architecture of spatio-temporal block (ST-block). (c) The architecture of the multi-branch temporal convolution.

forecasting, such as DCRNN (Li et al., 2017) and MST-GCN (Guo et al., 2019). Moreover, the performance of 8 spatio-temporal GNN models in forecasting different meteorological factors varies widely. For example, STGCN (Yu et al., 2017) ranks third in humidity forecasting, but worst in visibility forecasting. CLCRN (Lin et al., 2022) achieves the best performance because the conditional local kernel is specifically designed to capture the meteorological local spatial patterns. In addition, the current GNN baselines focus on the univariate forecasting and do not consider correlated multiple variables as cooperative inputs, which may have important impacts on weather forecasting.

Therefore, there are two urgent problems in the application of spatio-temporal forecasting in the meteorological domain: (1) A novel and general GNN framework designed specifically for the task of weather forecasting. (2) Make effective use of the multivariate meteorological factors.

5 MFMGCN

It is insufficient to simply use distance similarity to represent the correlations among observed data with spatio-temporal attributes (Geng et al., 2019). Modeling with a single graph usually brings unexpected bias, while multiple graphs can attenuate and offset the bias. To solve the above problems, we introduce MFMGCN, which combines several static graphs with the dynamic graph to effectively construct the intrinsic correlation among geographic locations based on meteorological factors. The framework of MFMGCN is shown in Figure 2. Through extensive experiments and ablation studies, we have demonstrated that MFMGCN achieves considerable forecasting performance gain and shows great robustness in the temporal dimension compared with previous methods on Weather2K-R.

5.1 Problem Setting

In this section, we describe the problem setting of the spatio-temporal weather forecasting. The meteorological network can be represented by a graph $G = (V, E, \mathbf{A})$, where V denotes the set of locations, E is a set of edges indicating the connectivity between the nodes and $\mathbf{A} \in \mathbf{R}^{N\times N}$is the adjacency matrix. Let $\mathbf{X}^{(t)} \in \mathbf{R}^{N\times D}$ represent the value of meteorological factors in all locations at the t-th interval,where N and D is the number of locations and factors, respectively. The spatio-temporal weather forecasting problem is to learn a function P that maps historical $W^{'}$ observed signals to the future W signals:

$$\left[\mathbf{X}^{(t-W'+1)}, \ldots, \mathbf{X}^{(t)}; G\right] \xrightarrow{P} \left[\mathbf{X}^{(t+1)}, \ldots, \mathbf{X}^{(t+W)}; G\right] \quad (1)$$

We set D = 1 for univariate forecasting in this study. We set $W^{'}$ = W = 12, which means both the input time length and forecasting time length are 12 steps.

5.2 Static Graph Construction

In this section, we describe in detail 4 static graphs used in MFMGCN that represent different types of correlations among geographic locations, including: (1) The distance graph $G_D = (V, E, \mathbf{A}_D)$. (2) The neighbor graph $G_N = (V, E, \mathbf{A}_N)$. (3) The pattern similarity graph $G_P = (V, E, \mathbf{A}_P)$. (4) The learnable graph $G_L = (V, E, \mathbf{A}_L)$.

5.2.1 Distance Graph

The distance-based graph describes the topological structure of the ground weather station network. Our distance graph G_D is constructed based on the spherical distance of stations' spatial location with thresholded Gaussian kernel (Shuman et al., 2013). The element of distance matrix $\mathbf{A}_D$ is defined as follows:

$$A_{D,ij} = \begin{cases} \exp\left(-\frac{d_{ij}^2}{\sigma_D^2}\right) & \text{, for } i \neq j \text{ and } \exp\left(-\frac{d_{ij}^2}{\sigma_D^2}\right) \geq \varepsilon \\ 0 & \text{, otherwise} \end{cases} \quad (2)$$

where d_{ij} represents the spherical distance between v_i and v_j. ε and σ_D^2 are used to control the sparsity and distribution of $\mathbf{A}_D$, respectively.

5.2.2 Neighbor Graph

We construct the neighbor graph G_N by simply connecting a location to its adjacent N_A stations. The element of the neighbor matrix $\mathbf{A}_N$ is defined as follows:

$$A_{N,ij} = \begin{cases} 1 & , v_i \text{ and } v_j \text{ are adjacent} \\ 0 & , \text{otherwise} \end{cases} \tag{3}$$

The influence of the selection value of N_A is discussed in Supplementary Material D.5.

5.2.3 Pattern Similarity Graph

Since distant stations may also have highly consistent meteorological characteristics, it is essential to establish a graph by mining pattern similarity between nodes based on our multivariate dataset. The element of pattern similarity matrix $\mathbf{A}_P$ is defined as follows:

$$A^f_{P,ij} = \begin{cases} \frac{\sum_{p=1}^{P}\left(t^f_{i,p}-\overline{T^f_i}\right)\left(t^f_{j,p}-\overline{T^f_j}\right)}{\sqrt{\sum_{i=1}^{p}\left(t^f_{i,p}-\overline{T^f_i}\right)^2}\sqrt{\sum_{j=1}^{p}\left(t^f_{j,p}-\overline{T^f_j}\right)^2}} & , \text{if } i \neq j \\ 0 & , \text{otherwise} \end{cases} \tag{4}$$

Note that the pattern similarity graph can be constructed using Pearson correlation coefficients (Zhang et al., 2020) based on any meteorological factor $f \in F$, where F denotes the set of all factors in Weather2K. For a specific feature f, $T^f_i = \left\{t^f_{i,1}, t^f_{i,2}, \cdots, t^f_{i,p}, \cdots, t^f_{i,P}\right\}$ describe the time-series sequence of v_i used for training, where P is the length of the series, and $t^f_{i,p}$ is the time-series data value of the vertex v_i at time step p. To model the correlation between any geographic location comprehensively, we build pattern graphs on temperature, visibility, and humidity.

5.2.4 Learnable Graph

The meteorological data need not only the accumulation of long time series for historical vertical comparative study, but also the horizontal comparative study of global regions. Despite the above graphs can capture spatial dependencies in detail, the information of the node itself is not sufficiently utilized due to the lack of historical level and geographic information which is challenging to predefine. Inspired by Wu et al. (2020b), we propose static node embedding $\mathbf{E} \in \mathbf{R}^D$ which is learnable to encode meteorological statistics for each node and compute:

$$\mathbf{M}_i = \tanh\left(\alpha \mathbf{E}_i \mathbf{\Theta}_1\right) \tag{5}$$

$$\mathbf{M}_j = \tanh\left(\alpha \mathbf{E}_j \mathbf{\Theta}_2\right) \tag{6}$$

$$A_{L,ij} = \mathrm{ReLU}\left(\tanh\left(\alpha\left(\mathbf{M}_i \cdot \mathbf{M}_j - \mathbf{M}_j \cdot \mathbf{M}_i\right)\right)\right) \tag{7}$$

where $\mathbf{E}_i$ and $\mathbf{E}_j$ represents randomly initialized node embeddings of node v_i and v_j, respectively. $\mathbf{\Theta}_1$,$\mathbf{\Theta}_2$ are linear layers and α is a hyper-parameter for controlling the saturation rate of the activation function. To make G_L more applicable to the meteorological field, each node of G_L has more connections with other nodes in different geographic locations, rather than selecting only the top-k closest nodes as its connections. And the edges of G_L are not limited to uni-directional because the geographical situation has two-way effects under different conditions.

5.3 Dynamic Graph Construction

In this section, the construction of the dynamic graph is presented. A node may have different types and degrees of weather, leading to dynamic changes in the relationship with other nodes. However, it is still hard to model such dynamic characteristics by static adjacency matrix based on distance or neighbor. Attention mechanism methods make computation and memory grow quadratically with the increase of graph size while RNN-based methods are limited by the length of the input sequence. Both of them need to be carefully modified to ensure successful training on Weather2K-R. We dynamically construct a graph $G_K = (V, E, \mathbf{A}_K)$ to model nonlinear spatio-temporal correlations in a efficient way.

Note that $\mathbf{X}_{t,i} = \{x_{t-W+1,i}, \cdots, x_{t-1,i}, \cdots, x_{t,i}\} \in \mathbf{R}^{W \times D}$ is an input temporal sequence of vertex v_i at time t, where W is the sequence length and D is the number of factors. We then simply flatten the vector to $\mathbf{Z}_i \in \mathbf{R}^{WD}$ to represent the characteristics on the input temporal scale. Similar to the way of building the learnable graph, the element of matrix $\mathbf{A}_K$ is defined as follows:

$$\mathbf{D}_i = \tanh\left(\beta \mathbf{Z}_i \mathbf{W}_1\right) \tag{8}$$

$$\mathbf{D}_j = \tanh\left(\beta \mathbf{Z}_j \mathbf{W}_2\right) \tag{9}$$

$$A_{K,ij} = \mathrm{ReLU}\left(\tanh\left(\beta\left(\mathbf{D}_i \cdot \mathbf{D}_j - \mathbf{D}_j \cdot \mathbf{D}_i\right)\right)\right) \tag{10}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are linear layers and β is a hyper-parameter for controlling the saturation rate of the activation function. Different from the learnable graph based on static node embedding, G_K is generated dynamically both in the training and inference stages.

5.4 Multi-graph Fusion

Since different graphs contain specific spatio-temporal information and make unequal contributions to the forecasting result of every single node, it is inappropriate to simply merge them using weighted sum or other averaging approaches in a unified way. Therefore, we use learnable weights $\mathbf{W}_s \in \mathbf{R}^{N \times N}$ to describe the importance of graph s for each node, where N presents the number of nodes and $s \in S = \{D, N, P, L, K\}$ is one of the proposed graphs. Finally, the result of multi-graph fusion is $G_{fused} = (V, E, \mathbf{A}_{fused})$:

$$\mathbf{A}_{fused} = \sum_{s \in S} \mathbf{W}_s \odot \mathbf{A}_s \tag{11}$$

where $\odot$ indicates the element-wise Hadamard product.

5.5 Spatio-Temporal Graph Neural Network

Considering the time-space consistency is important in weather forecasting, we build a model by stacking spatio-temporal blocks (ST-block) which are complete convolutional structures. As shown in Figure 2 (b), the ST-block adopts a residual learning framework which consists of a spatial convolution and a temporal convolution in order. Then, a fully-connected output layer is used to generate the final prediction $\hat{\mathbf{Y}}$.

5.5.1 Graph Convolution in Spatial Dimension

Following Yu et al. (2017), we adopt the spectral graph convolution to directly process the signals at each time slice. In spectral graph analysis, taking the weather conditions of factor f at time t as an example, the signal all over the graph is $x = \mathbf{x}_t^f \in \mathbf{R}^N$, then the signal on the graph G is filtered by a kernel g_θ:

$$g_\theta *_G x = g_\theta(\mathbf{L})x = g_\theta\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\right)x = \mathbf{U}g_\theta(\mathbf{\Lambda})\mathbf{U}^T x \quad (12)$$

where graph Fourier basis $\mathbf{U} \in \mathbf{R}^{N\times N}$ is the matrix of eigenvectors of the normalized graph Laplacian $\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \in \mathbf{R}^{N\times N}$. Note that $\mathbf{I}_N$ is an identity matrix and $\mathbf{D} \in \mathbf{R}^{N\times N}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \Sigma_j A_{ij}$ ($\mathbf{A}$ is the adjacent matrix). $\mathbf{\Lambda} \in \mathbf{R}^{N\times N}$ is the diagonal matrix of eigenvalues of $\mathbf{L}$ and filter $g_\theta(\mathbf{\Lambda})$ is also a diagonal matrix. The above formula can be understood as Fourier transforming g_θ and x respectively into the spectral domain, and multiplying their transformed results, and doing the inverse Fourier transform to get the output of the graph convolution (Shuman et al., 2013).

In practice, Chebyshev polynomials are adopted to solve the efficiency problem of performing the eigenvalue decomposition since the scale of our graph is large (Hammond et al., 2011):

$$g_\theta *_G x = g_\theta(\mathbf{L})x = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\mathbf{L}})x \quad (13)$$

where $T_k(\tilde{\mathbf{L}}) \in \mathbf{R}^{N\times N}$ is the Chebyshev polynomial of order k evaluated at the scaled Laplacian $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{\max} - \mathbf{I}_N$, and $\lambda_{\max}$ denotes the largest eigenvalue of $\mathbf{L}$. Note that K is the kernel size of graph convolution, which determines the maximum radius of the convolution from central nodes.

5.5.2 Convolution in Temporal Dimension

Although RNN-based methods have superiority over CNN-based methods in capturing temporal dependency, the time-consuming problem of recurrent networks is nonnegligible, especially in the field of long-term weather forecasting.

In MFMGCN, a multi-branch convolution layer on the time axis is further stacked to model the temporal dynamic behaviors of meteorological data. As shown in Figure 2 (c), the convolution layer of each branch has different receptive fields for extracting information at different scales. Information of different scales is then fused through concatenation and a 1x1 convolution layer.

5.5.3 Time-space Consistency

The time-space consistency in weather forecasting means that the larger the time scale of the data, the wider the spatial range that can be predicted. Inspired by this rule, we balance the span of space and time by setting appropriate parameters in ST-blocks. The kernel size K of graph convolution and receptive field in temporal convolution are increased linearly with the stacking of blocks. By doing so, our model can integrate spatio-temporal information orderly and avoid imbalance between both sides.

5.6 Experiments

5.6.1 Ablation Study

In this section, we conduct ablation studies to show the effects of different fusion selections and discuss whether each graph works as intended in forecasting temperature, visibility, and humidity. We first evaluate the performance of each graph separately. Secondly, we set distance graph G_D as the base input, which is the same as baseline models, to verify whether the stack of more graphs (i.e. more information) leads to better performance. And then we compare the performance of all graphs but one (i.e. 4-graph fusion). Finally, we show the performance of 5-graph fusion.

From Table 4, we can conclude that: (1) G_D and G_N is similar in results because they are both constructed based on location relationships, and it's better to combine the two

Table 4: The results of ablation studies.

G_D	G_N	G_P	G_L	G_K	Temperature		Visibility		Humidity	
					MAE	RMSE	MAE	RMSE	MAE	RMSE
✔					1.7190	2.5056	4.2625	6.3809	8.1178	11.6013
	✔				1.7420	2.5211	4.2724	6.4342	8.0778	11.6008
		✔			2.3651	3.1726	4.8104	6.5291	8.6860	12.2008
			✔		1.5842	2.2753	4.1551	6.1609	7.2361	10.4721
				✔	1.8267	2.5691	4.4877	6.3925	7.9877	11.5218
✔	✔				1.7430	2.5342	4.2403	6.3334	7.9565	11.4717
✔	✔	✔			1.6559	2.3834	4.1486	6.1512	7.7309	11.1712
	✔	✔	✔	✔	1.5540	2.2448	3.9409	6.0163	7.1181	10.3577
✔		✔	✔	✔	1.5266	2.1925	3.9374	6.0278	7.1326	10.4074
✔	✔		✔	✔	1.5046	2.1769	3.9361	6.0685	7.2506	10.5775
✔	✔	✔		✔	1.5893	2.2887	4.4932	6.4680	7.4027	10.7293
✔	✔	✔	✔		1.5053	2.1677	4.0174	6.0701	7.1858	10.4025
✔	✔	✔	✔	✔	**1.4418**	**2.0574**	**3.9277**	**6.0132**	**7.1125**	**10.3512**

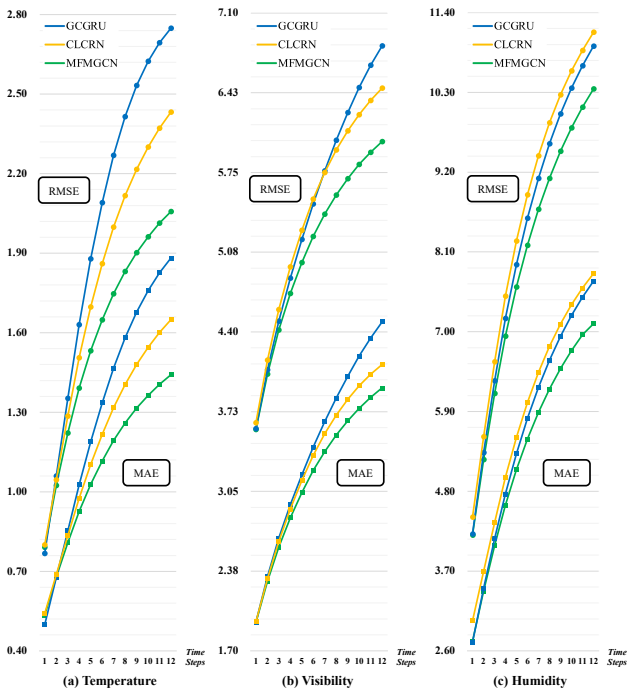

Figure 3: Overall method comparison in forecasting (a) temperature, (b) visibility, and (c) humidity.

graphs to make more efficient use of geographical distribution. (2) Using G_P separately yields a quite poor result. However, G_P brings a positive influence in multi-graph fusion. (3) G_L is the optimal single graph, and it also brings considerable performance gain in multi-graph fusion. (4) Using G_K separately does not work as well because it lacks the stable correlation provided by the static graph. But it can be complementary to the static graph and brings expected gain. (5) Whether forecasting temperature, visibility, or humidity, the effects of different fusion selections have similar trend on performance, which indicates that the multi-graph method is universal to meteorological factors.

5.6.2 Overall Method Comparison

To avoid complicated and verbose plots, we give the comparison of methods with the overall top three performances on MAE and RMSE: GCGRU (Seo et al., 2018), CLCRN (Lin et al., 2022), and our MFMGCN in Figure 3. With the increase of the forecasting horizon from the first time step to 12 time steps, the gap between MFMGCN and other methods gradually widened. It can be concluded that MFMGCN improves both the forecasting performance and temporal robustness, which is of great significance for real-world practical applications. In addition, MFMGCN achieves the state-of-the-art performance in different meteorological factors, which further verifies it is a general framework for the task of weather forecasting.

6 CONCLUSION

This paper contributes a new benchmark dataset named Weather2K, which aims to make up for the deficiencies of existing weather forecasting datasets in terms of real-time, reliability, and diversity. The data of Weather2K is hourly collected from 2,130 ground weather stations, which contain 20 meteorological factors and 3 constants for position information. Weather2K is applicable to diverse tasks such as time series forecasting and spatio-temporal forecasting. As far as we know, Weather2K is the first attempt to tackle weather forecasting task by taking full advantage of the strengths of observation data from ground weather stations. Besides, we further propose MFMGCN, a novel and general framework designed specifically for the weather forecasting task, to effectively construct the intrinsic correlation among geographic locations based on meteorological factors by combining the static graph and the dynamic graph, which significantly outperforms previous spatio-temporal GNN methods on Weather2K. We hope that the release of Weather2K can provide a foundation for accelerated research in this area and foster collaboration between atmospheric and data scientists.

Acknowledgements

This work is partially supported by the MoE-CMCC "Artifical Intelligence" Project (Grant No. MCM20190701) and the BUPT innovation and entrepreneurship support program (Grant No. 2023-YC-A177).

References

Jiandong Bai, Jiawei Zhu, Yujiao Song, Ling Zhao, Zhixiang Hou, Ronghua Du, and Haifeng Li. A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS International Journal of Geo-Information*, 10(7):485, 2021.

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in neural information processing systems*, 33:17804–17815, 2020.

Hylke E Beck, Albert IJM Van Dijk, Vincenzo Levizzani, Jaap Schellekens, Diego G Miralles, Brecht Martens, and Ad De Roo. Mswep: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1):589–615, 2017.

Marco L Carrera, Stéphane Bélair, and Bernard Bilodeau. The canadian land data assimilation system (caldas): Description and synthetic evaluation study. *Journal of Hydrometeorology*, 16(3):1293–1314, 2015.

Matteo De Felice, Andrea Alessandri, and Paolo M Ruti. Electricity demand forecasting over italy: Potential ben-

efits using numerical weather prediction models. *Electric Power Systems Research*, 104:71–79, 2013.

Michaël Defferrard, Martino Milani, Frédérick Gusset, and Nathanaël Perraudin. Deepsphere: a graph-based spherical cnn. *arXiv preprint arXiv:2012.15000*, 2020.

Imme Ebert-Uphoff, David R Thompson, Ibrahim Demir, Yulia R Gel, Anuj Karpatne, Mariana Guereque, Vipin Kumar, Enrique Cabral-Cano, and Padhraic Smyth. A vision for the development of benchmarks to bridge geoscience and data science. In *17th International Workshop on Climate Informatics*, 2017.

Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3656–3663, 2019.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 922–929, 2019.

David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2): 129–150, 2011.

Jie He, Kun Yang, Wenjun Tang, Hui Lu, Jun Qin, Yingying Chen, and Xin Li. The first high-resolution meteorological forcing dataset for land process studies over china. *Scientific Data*, 7(1):1–11, 2020.

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

George J Huffman, David T Bolvin, Dan Braithwaite, Kuolin Hsu, and Robert Joyce. Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg). 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Christian Kummerow, J Simpson, O Thiele, W Barnes, ATC Chang, E Stocker, RF Adler, A Hou, R Kakar, F Wentz, et al. The status of the tropical rainfall measuring mission (trmm) after two years in orbit. *Journal of applied meteorology*, 39(12):1965–1982, 2000.

Jia Li, Zhichao Han, Hong Cheng, Jiao Su, Pengyun Wang, Jianfeng Zhang, and Lujia Pan. Predicting path failure in time-evolving graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1279–1289, 2019a.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019b.

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

Haitao Lin, Zhangyang Gao, Yongjie Xu, Lirong Wu, Ling Li, and Stan Z Li. Conditional local convolution for spatio-temporal meteorological forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7470–7478, 2022.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*, 2021.

Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.

Eike H Müller and Robert Scheichl. Massively parallel solvers for elliptic partial differential equations in numerical weather and climate prediction. *Quarterly Journal of the Royal Meteorological Society*, 140(685):2608–2624, 2014.

Andreas Holm Nielsen, Alexandros Iosifidis, and Henrik Karstoft. Cloudcast: A satellite-based dataset and baseline for forecasting clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3485–3494, 2021.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4838–4845, 2021.

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

Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems*, 30, 2017.

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

Matthew Rodell, PR Houser, UEA Jambor, J Gottschalck, Kieran Mitchell, C-J Meng, Kristi Arsenault, B Cosgrove, J Radakovich, M Bosilovich, et al. The global land data assimilation system. *Bulletin of the American Meteorological society*, 85(3):381–394, 2004.

Benedek Rozemberczki, Paul Scherer, Yixuan He, George Panagopoulos, Alexander Riedel, Maria Astefanoaei, Oliver Kiss, Ferenc Beres, Guzmán López, Nicolas Collignon, et al. Pytorch geometric temporal: Spatiotemporal signal processing with neural machine learning models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4564–4573, 2021.

Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *International conference on neural information processing*, pages 362–373. Springer, 2018.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3): 83–98, 2013.

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

Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.

MA Tolstykh and AV Frolov. Some current problems in numerical weather prediction. *Izvestiya Atmospheric and Oceanic Physics*, 41(3):285–295, 2005.

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

Vladimir Vovk. Kernel ridge regression. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116, 2013.

Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2. 5-gnn: A domain knowledge enhanced graph neural network for pm2. 5 forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 163–166, 2020.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430, 2021.

Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, and Junzhou Huang. Adversarial sparse transformer for time series forecasting. *Advances in neural information processing systems*, 33:17105–17115, 2020a.

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 *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763, 2020b.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Xu Zhang, Ruixu Cao, Zuyu Zhang, and Ying Xia. Crowd flow forecasting with multi-graph neural networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21 (9):3848–3858, 2019.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1234–1241, 2020.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. *arXiv preprint arXiv:2201.12740*, 2022.

Supplementary Materials: Weather2K: A Multivariate Spatio-Temporal Benchmark Dataset for Meteorological Forecasting Based on Real-Time Observation Data from Ground Weather Stations

A NOTATION DEFINITION

Table 5 shows the glossary of notations used in this paper.

Table 5: Glossary of notations.

Symbol	Meaning
$G = (V, E, \mathbf{A})$	Graph represented nodes, edges and adjacency matrix respectively.
v_i	The i-th node.
N	The number of geographic locations.
D	The number of used factors.
$\mathbf{X}^{(t)}$	The value of meteorological factors in all locations at the t-th interval.
$W^{'}$	The input time length.
W	The forecasting time length.
F	The set of all factors in Weather2K.
f	A specific factor in F.
$t_{i,p}^{f}$	The time-series data value of the vertex v_i at time step p of factor f.
T_i^f	The time-series sequence of v_i used for training of factor f.
α, β	A hyper-parameter for controlling the saturation rate of the activation function.
$\mathbf{\Theta}$	The linear layer.
$\mathbf{E}_i$	The static node embedding of v_i.
$\mathbf{Z}_i$	The dynamic node embedding of v_i.
S	The set of proposed graphs.
s	A specific graph in S.
W_s	Learnable weights which describe the importance of graph s for each node.
$\mathbf{Y}^{(t,W)}$	The ground truth at time t.
$\hat{\mathbf{Y}}^{(t,W)}$	The final prediction at time t.
$\mathbf{x}_t^f$	The signal of factor f at time t all over the graph.
g_θ	A kernel to filter the signal.
$\mathbf{L}$	The normalized graph Laplacian.
$\mathbf{U}$	The matrix of eigenvectors of $\mathbf{L}$.
$\mathbf{I}_N$	An identity matrix.
$\mathbf{D}$	The diagonal degree matrix.
$\mathbf{\Lambda}$	The diagonal matrix of eigenvalues of $\mathbf{L}$
$\tilde{\mathbf{L}}$	The scaled Laplacian.
$T_k(\tilde{\mathbf{L}})$	The Chebyshev polynomial of order k evaluated at the scaled Laplacian $\tilde{\mathbf{L}}$.

B METRICS COMPUTATION

For time series forecasting, let $\mathbf{T}_i \in \mathbf{R}^{L\times C}$ be the ground truth, and $\hat{\mathbf{T}}_i$ be the predictions given by neural networks, where L denotes the forecasting time length and C is the number of forecasting factors. The two metrics including MAE, MSE are calculated as:

$$\mathrm{MAE}\left(\mathbf{T}_i, \hat{\mathbf{T}}_i\right) = \frac{1}{LC}\left|\hat{\mathbf{T}}_i - \mathbf{T}_i\right| \tag{14}$$

$$\mathrm{MSE}\left(\mathbf{T}_i, \hat{\mathbf{T}}_i\right) = \frac{1}{LC}\left(\hat{\mathbf{T}}_i - \mathbf{T}_i\right)^2 \tag{15}$$

For spatio-temporal weather forecasting, let $\mathbf{Y}^{(t,W)} = \left[\mathbf{X}^{(t+1)}, \ldots, \mathbf{X}^{(t+W)}\right]$ be the ground truth, and $\hat{\mathbf{Y}}^{(t,W)} = \left[\hat{\mathbf{X}}^{(t+1)}, \ldots, \hat{\mathbf{X}}^{(t+W)}\right]$ be the predictions given by graph models. The two metrics including MAE, RMSE are calculated as:

$$\mathrm{MAE}\left(\mathbf{Y}^{(t,W)}, \hat{\mathbf{Y}}^{(t,W)}\right) = \frac{1}{WND}\sum_{i=t+1}^{t+W}\left|\hat{\mathbf{X}}^{(i)} - \mathbf{X}^{(i)}\right| \tag{16}$$

$$\mathrm{RMSE}\left(\mathbf{Y}^{(t,W)}, \hat{\mathbf{Y}}^{(t,W)}\right) = \sqrt{\frac{1}{WND}\sum_{i=t+1}^{t+W}\left(\hat{\mathbf{X}}^{(i)} - \mathbf{X}^{(i)}\right)^2} \tag{17}$$

where D, N is the number of forecasting factors and geographic locations respectively, W is the forecasting time length.

C MORE DETAILS ON THE DATASET OF WEATHER2K

C.1 The Data Processing

For sensor noise, bias and external effects when collecting data, CMA's QC technologies mainly include numerical range check, climatic range check, main change range check, internal consistency check, temporal consistency check, spatial consistency check. Moreover, there are six main steps in our procedure of data processing, which are listed as follows:

(1) Acquisition of the Raw Data The raw data comes from CMA's observation data of China's ground weather stations. The hourly updated data contains a total of 105 variables from approximately 2,500 ground weather stations covering an area of 6 million square kilometers. We select the time period from January 2017 to August 2021, in which we can ensure the use of unified meteorological observation instruments and sensors for data collection.

(2) Screening of the Meteorological Factors Based on the consideration of building a representative and comprehensive dataset for meteorological forecasting, we select 3 time-invariant constants: latitude, longitude, and altitude to provide position information and 20 important meteorological factors such as air pressure, air temperature, dewpoint temperature, water vapor pressure, relative humidity, wind speed, wind direction, vertical visibility, horizontal visibility, precipitation, etc. The definitions and physical descriptions of the chosen variables are listed in Table 1.

(3) Screening of the Time Series In this step, we mainly ensure the integrity of the time series of each ground weather station. According to the chosen time range, the length of the time series should be 40,896 steps for a single ground weather station. We discard the whole data of a single weather station with a missing proportion above the upper limit, which is set to 1% of the entire time steps. For the remaining data, we use linear interpolation. After this step, the number of eligible ground weather stations drops to 2,373.

(4) Screening of the Default Values There are default values for meteorological factors represented by certain numbers, such as 999,999 for visibility. For a single ground weather station, if the default value proportion for any one of the 20 meteorological factors exceeds 1%, we will discard the entire data. We also use linear interpolation for the default values in the remaining data. After this step, the number of eligible ground weather stations falls further to 2,130.

(5) Manual Tuning of the Outliers Despite CMA's QC technologies for the data, we found some unexpected outliers. We use box plots to make statistics for each meteorological factor. We focus on discrete individual points in the box plot that are above the upper whisker or below the lower whisker. The opinions of 3 experts in the field of meteorological science help us decide whether these points are outliers for recording errors or acceptable extremes.

(6) Storage of the Data We use a huge Numpy format file to store the processed data. The historical data of each individual ground weather station constitutes the time series of multivariate meteorological factors. Thousands of stations form the spatial sequences, each of which is assigned a fixed index number based on the original station number. In the release version of the dataset, we give the index correspondence in a text file.

C.2 Dataset Statistics

C.2.1 Statistics of Meteorological Factors by Box Plot

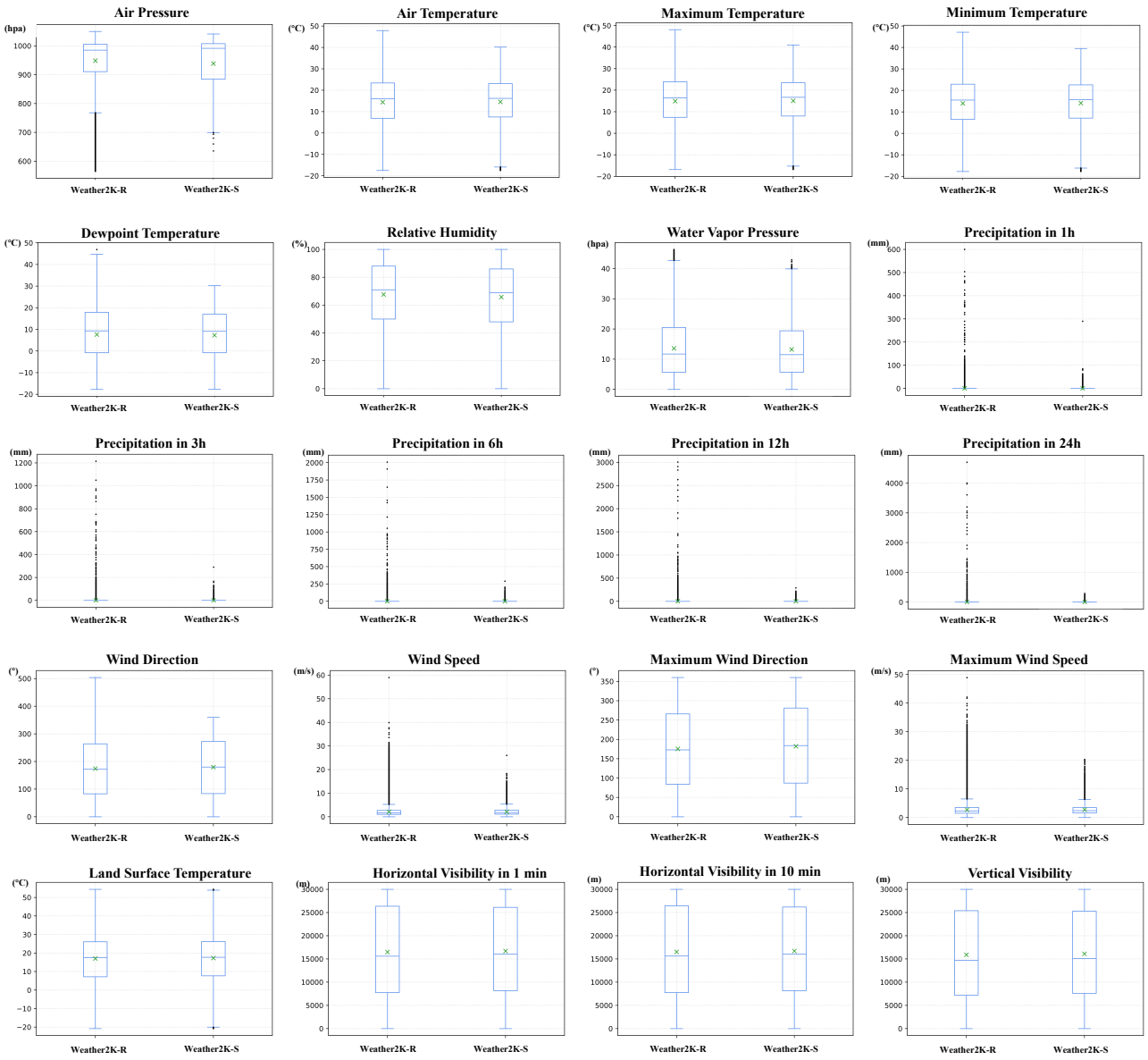

Figure 4: Box plots of 20 meteorological factors in Weather2K-R and Weather2K-S.

In descriptive statistics, the box plot is a method for graphically reflecting the overall characteristics of numerical data distribution through their quartiles. Although we cannot apply a simple Gaussian distribution for modelling various weather properties, such as precipitation and wind, it can also play an auxiliary role in outlier processing. Figure 4 shows the box plots of 20 meteorological factors in Weather2K-R and Weather2K-S. The top and bottom boundaries of a box are the upper and lower quartiles of the statistic indices at these stations, while the line and the cross inside the box are the median and the mean value, respectively. The vertical dashed lines extending from the box represent the minimum and maximum of the corresponding indices. "Outliers" (acceptable after expert discussion) are plotted as individual black points.

C.2.2 Statistics of Weather2K-S by Cumulative Distribution Function

In probability theory and statistics, the Cumulative Distribution Function (CDF) can completely describes the probability distribution of a real-valued variable. Figure 5 shows the CDF of 20 meteorological factors in Weather2K-S. Lines 1-15 denote the different ground weather stations in Weather2K-S, which can be intuitively found that the stations we select are representative and diverse.

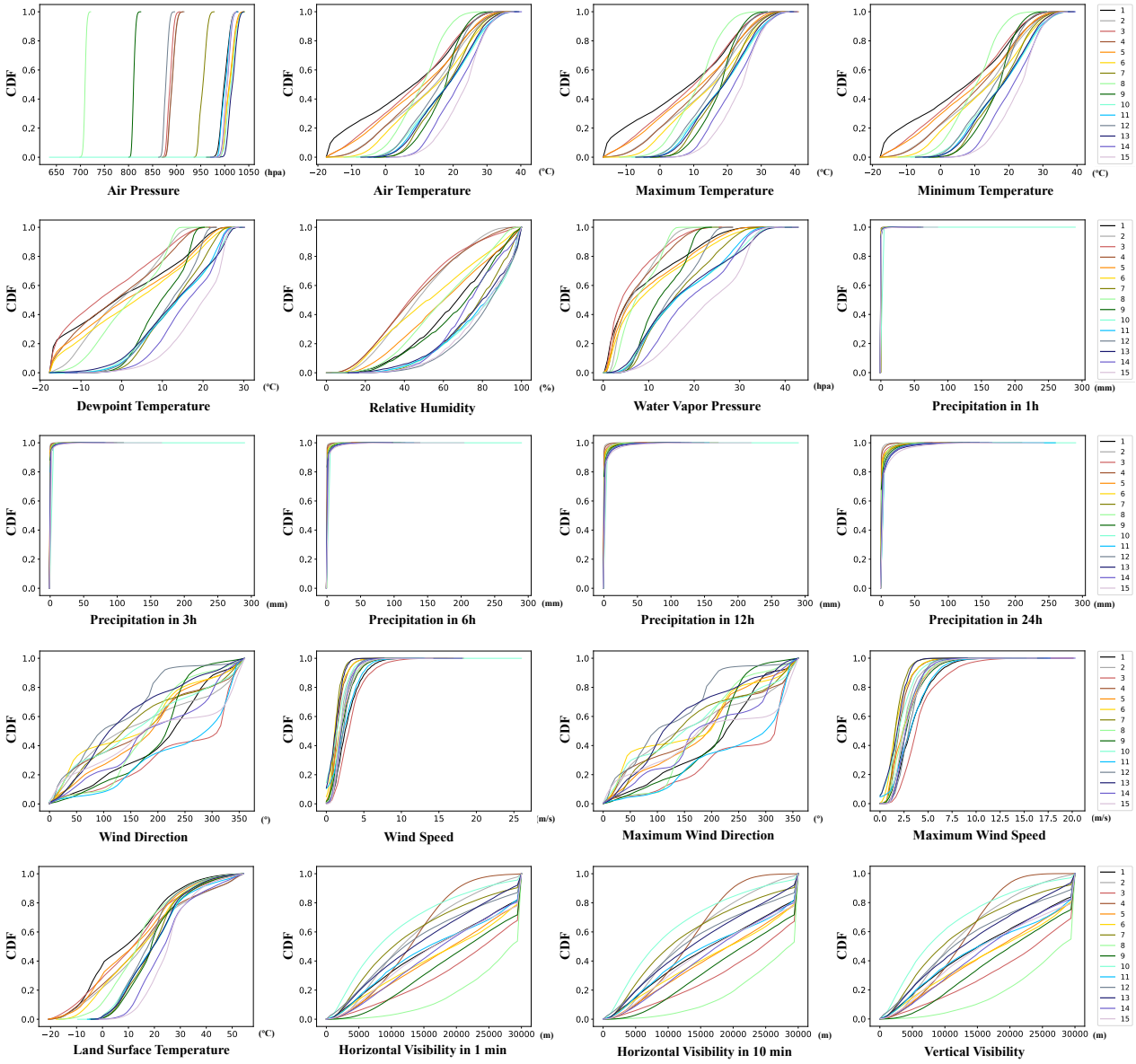

Figure 5: CDF of 20 meteorological factors in Weather2K-S.

C.2.3 Statistics of Weather2K-R by Administrative Region

Figure 6 shows the statistics of ground weather stations of each province in Weather2K-R to facilitate the potential use of specific audiences. Weather2K-R covers 31 provinces in China and contains 1,866 ground weather stations, with an average of more than 60 ground weather stations per province. Among them, Hebei has the largest number of 132 stations, while Shanghai and Tianjin have the smallest number of 10 stations.

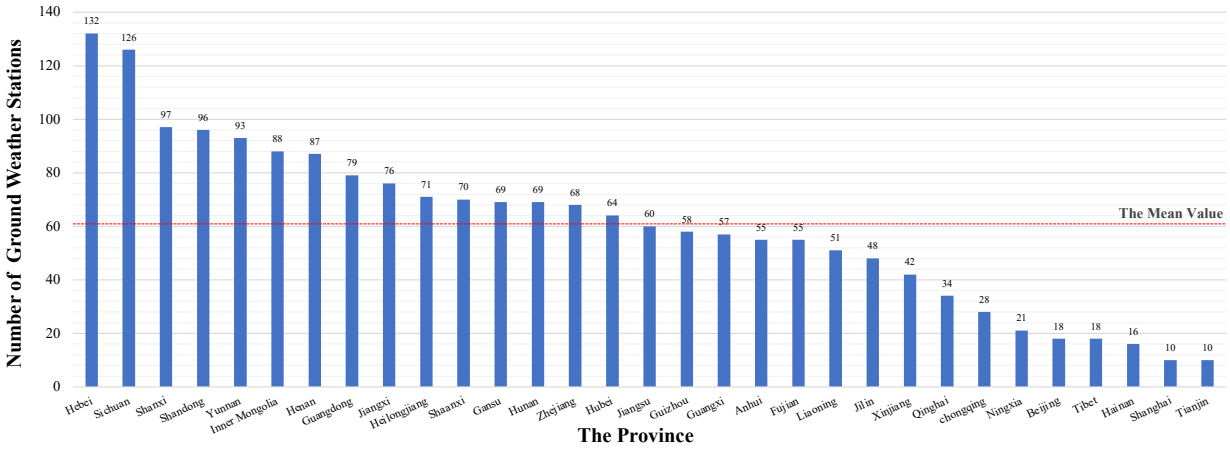

Figure 6: The statistics of ground weather stations of each province in Weather2K-R.

D EXPERIMENT

D.1 Baseline Models

In the application direction of time series forecasting, four representative transformer-based baselines are set up for comparison on Weather2K-S.

Transformer Vaswani et al. (2017) propose a simple network architecture, the Transformer, based on the self-attention mechanism. Transformers show great modeling ability for long-range dependencies and interactions in sequential data and thus are attractive in time series modeling. However, it is computationally prohibitive to applying self-attention to long-term time series forecasting due to the quadratic complexity of sequence length L in both memory and time. Many Transformer variants have been proposed to adapt to long-term time series forecasting in a efficient way.

Reformer Kitaev et al. (2020) introduce two techniques to improve the efficiency of Transformers. For one, Reformer replaces dot-product attention by one that uses locality-sensitive hashing, changing its complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L \log L)$. Furthermore, Reformer uses reversible residual layers instead of the standard residuals, which allows storing activations only once in the training process instead of N times, where N is the number of layers.

Informer Zhou et al. (2021) propose ProbSparse self-attention mechanism to efficiently replace the canonical self-attention, which use the Kullback-Leibler divergence distribution measurement between queries and keys to select $\mathcal{O}(\log L)$ dominant queries. Informer achieves the $\mathcal{O}(L \log L)$ time complexity and memory usage on dependency alignments. It also designs a generative style decoder to produce long sequence output with only one forward step to avoid accumulation error.

Autoformer Wu et al. (2021) devise a seasonal-trend decomposition architecture with an auto-correlation mechanism working as an attention module. The auto-correlation block measures the time-delay similarity between inputs signal and aggregate the top-k similar sub-series to produce the output. Autoformer achieves $\mathcal{O}(L \log L)$ complexity and breaks the information utilization bottleneck by expanding the point-wise representation aggregation to sub-series level.

In the application direction of spatio-temporal forecasting, eight state-of-the-art spatio-temporal GNN models are set up for comparison on Weather2K-R.

STGCN Yu et al. (2017) propose a deep learning framework for traffic prediction, integrating graph convolution and gated

temporal convolution through spatio-temporal convolutional blocks. The model is entirely composed of convolutional structures and therefore achieves parallelization over input with fewer parameters and faster training speed.

DCRNN Li et al. (2017) study the traffic forecasting problem and model the spatial dependency of traffic as a diffusion process on a directed graph. DCRNN captures both spatial and temporal dependencies among time series using diffusion convolution and the sequence to sequence learning framework together with scheduled sampling.

GCGRU Seo et al. (2018) combine convolutional neural networks on graphs to identify spatial structures and recurrent neural networks to find dynamic patterns. GCGRU shows that exploiting simultaneously graph spatial and dynamic information about data can improve both precision and learning speed.

ASTGCN Guo et al. (2019) develop a attention mechanism to learn the dynamic spatio-temporal correlations of traffic data. The spatio-temporal convolution module consists of graph convolutions for capturing spatial features from the traffic network structure and convolutions in the temporal dimension for describing dependencies from nearby time slices.

MSTGCN Guo et al. (2019) propose a degraded version of ASTGCN, which gets rid of the spatio-temporal attention.

TGCN Zhao et al. (2019) propose the TGCN model by combining the graph convolutional network and gated recurrent unit, which are used to capture the topological structure of the road network to model spatial dependence and the dynamic change of traffic data on the roads to model temporal dependence, respectively.

A3TGCN Bai et al. (2021) propose the A3TGCN model to simultaneously capture global temporal dynamics and spatial correlations. The attention mechanism was introduced to adjust the importance of different time points and assemble global temporal information to improve prediction accuracy.

CLCRN Based on the assumption of smoothness of location-characterized patterns, Lin et al. (2022) propose the conditional local kernel and embed it in a graph-convolution-based recurrent network to model the temporal dynamics for spatio-temporal meteorological forecasting.

D.2 Implementation Details

All the deep learning networks mentioned in this paper are implemented in PyTorch (Paszke et al., 2019) of version 1.8.0 with CUDA version 11.1. For the hyper-parameters that are not specifically mentioned in each model, we use the default settings from their original proposals. Every single experiment is executed on a server with 4 NVIDIA GeForce RTX3090 GPUs with 24GB of video memory.

D.2.1 Transformers in Time Series Forecasting

Experiments are carried out on Weather2K-S with the temporal scale from January 1, 2019 to December 31, 2020. The data is divided into the training set, validation set, and test set according to the ratio of 3:1:2. The batch size is set to 32. The global seed is set to 2021 for the experiment repeat. The train epoch is set to 30, while the training process is early stopped within 5 epochs if no loss degradation on the validation set is observed. All the models are trained with L2 loss, using the Adam (Kingma and Ba, 2014) optimizer with the learning rate of $1e^{-4}$.

D.2.2 GNNs in Spatio-Temporal Forecasting

We truncate Weather2K-R with the temporal scale from January 1, 2017 to December 31, 2019. The baselines are all implemented according to PyTorch Geometric Temporal (Rozemberczki et al., 2021). We construct a distance graph as one of the GNN models' inputs by utilizing three time-invariant constants in Weather2K-R that provide position information including latitude, longitude, and altitude. The data for training, validation, and testing are all non-overlapping one-year time, while the random seed is set to 2022. The batch size for training is set to 32. All the models are trained with the target function of MAE and optimized by Adam optimizer (Kingma and Ba, 2014) for 100 epochs. Early-stopping epoch is set to 50 according to the validation set. The initial learning rate is set to $1e^{-2}$, and it decays with the ratio $5e^{-2}$ per 10 epochs in the first 50 epochs.

D.3 The Results of Multivariate to Univariate Forecasting with Transformer Baselines

To take full advantage of multivariate data in Weather2K-S, we explore and attempt to use the multivariate data to forecast the target univariate meteorological factor. We conduct experiments on the assumption that there are some underlying

correlations between different meteorological factors. Without any prior knowledge, we use all 20 dimensional meteorological factors as input data. Table 6 shows the performance of multivariate to univariate forecasting. The number in the parenthesis is the difference value from the univariate result, where the better result is highlighted in bold. However, multivariate to univariate forecasting does not result in stable performance gain and even brought performance degradation in most cases compared to univariate forecasting. It can be concluded that multivariate information of meteorological factors is underutilized in existing transformer-based time series forecasting baselines.

Table 6: Multivariate to univariate results with different prediction lengths of 24, 72, 168, 336, and 720 on Weather2K-S. The input time length is 72 steps. The better results compared to univariate results are highlighted in **bold**.

Models		Transformer		Reformer		Informer		Autoformer	
Factors	**Metrics**	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Temperature	24	0.0874 (+0.0134)	0.2256 (+0.0256)	0.0820 **(-0.0077)**	0.2173 **(-0.0133)**	0.0953 (+0.0195)	0.2322 (+0.0291)	0.1114 (+0.0136)	0.2539 (+0.0172)
	72	0.1379 (+0.0133)	0.2843 (+0.0187)	0.1200 **(-0.0094)**	0.2669 **(-0.0128)**	0.1575 (+0.0303)	0.3033 (+0.0322)	0.1505 (+0.0121)	0.2968 (+0.0136)
	168	0.1624 (+0.0064)	0.3090 (+0.0091)	0.1354 **(-0.0077)**	0.2841 **(-0.0085)**	0.2002 (+0.0303)	0.3450 (+0.0288)	0.1642 (+0.0068)	0.3078 (+0.0076)
	336	0.1757 (+0.0048)	0.3224 (+0.0076)	0.1398 **(-0.0053)**	0.2886 **(-0.0061)**	0.2298 (+0.0180)	0.3785 (+0.0199)	0.1734 (+0.0079)	0.3180 (+0.0088)
	720	0.1939 (+0.0336)	0.3399 (+0.0341)	0.1468 **(-0.0013)**	0.2940 **(-0.0028)**	0.2550 (+0.0431)	0.3978 (+0.0341)	0.2037 (+0.0026)	0.3486 (+0.0033)
Visibility	24	0.7300 (+0.0195)	0.6696 (+0.0160)	0.6355 **(-0.0268)**	0.6101 **(-0.0187)**	0.7316 (+0.0309)	0.6555 (+0.0165)	0.8596 (+0.0582)	0.7188 (+0.0259)
	72	0.8943 (+0.0221)	0.7548 (+0.0073)	0.8072 **(-0.0232)**	0.7219 **(-0.0221)**	0.9388 (+0.0541)	0.7587 (+0.0084)	1.0087 (+0.0737)	0.7838 (+0.0253)
	168	0.9101 **(-0.0052)**	0.7721 **(-0.0081)**	0.9067 **(-0.0016)**	0.7672 **(-0.0123)**	0.9853 (+0.0425)	0.8014 (+0.0071)	1.0108 (+0.0197)	0.7873 (+0.0039)
	336	0.9883 **(-0.0069)**	0.7989 **(-0.0070)**	0.9203 **(-0.0016)**	0.7787 **(-0.0023)**	1.0229 **(-0.0003)**	0.8304 **(-0.0084)**	1.0135 (+0.0216)	0.7907 (+0.0075)
	720	1.0143 (+0.0180)	0.8117 (+0.0027)	0.9453 **(-0.0053)**	0.7964 (+0.0018)	1.0736 (+0.0501)	0.8539 (+0.0101)	1.0596 **(-0.0055)**	0.8092 **(-0.0053)**
Humidity	24	0.4307 (+0.0561)	0.5010 (+0.0403)	0.3632 (+0.0014)	0.4551 (+0.0022)	0.4318 (+0.0569)	0.5024 (+0.0461)	0.5217 (+0.0926)	0.5539 (+0.0598)
	72	0.5741 (+0.0513)	0.5863 (+0.0260)	0.4941 (+0.0009)	0.5454 **(-0.0001)**	0.5989 (+0.0757)	0.5985 (+0.0366)	0.6215 (+0.0704)	0.6140 (+0.0419)
	168	0.6427 (+0.0492)	0.6242 (+0.0238)	0.5457 (+0.0036)	0.5767 (+0.0008)	0.7281 (+0.1060)	0.6617 (+0.0454)	0.6536 (+0.0425)	0.6283 (+0.0241)
	336	0.6620 (+0.0513)	0.6354 (+0.0212)	0.5856 (+0.0095)	0.5992 (+0.0054)	0.8020 (+0.1055)	0.7129 (+0.0476)	0.6896 (+0.0382)	0.6456 (+0.0198)
	720	0.6668 (+0.0491)	0.6359 (+0.0213)	0.5819 **(-0.0045)**	0.5963 **(-0.0046)**	0.8803 (+0.0824)	0.7711 (+0.0426)	0.7358 (+0.0213)	0.6719 (+0.0120)

D.4 The Univariate Results of the Persistent Model and Classical Nonparametric Methods

In order to perform a meaningful comparison for the forecasting results, the persistent model should be introduced to quantify the improvement given by advanced forecasting techniques. The persistent model is the most cost-effective forecasting model (De Felice et al., 2013) which assumes that the conditions will not change, which means "today's weather is tomorrow's forecast". So it is often called the naive forecast. In addition, we present the univariate forecasting results of 3 classical nonparametric methods including linear regression (Su et al., 2012), ridge regression (McDonald, 2009), and kernel ridge regression (Vovk, 2013) on Weather2K-S in Table 7. Different from transformer baselines, we use all the data of Weather2K-S, with a 3:1 ratio of the training set to the test set.

Table 7: Univariate results of the persistent model and nonparametric methods with different prediction lengths of 24, 72, 168, 336, and 720 steps. The input time length is 72 steps. The reported results of mean and standard deviation are obtained through experiments on Weather2K-S. Results with **bold** and underlines are the best and worst performance, respectively.

Models		Persistent Model		Linear Regression		Ridge Regression		Kernel Ridge Regression	
Factors	**Metrics**	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Temperature	24	0.2610±0.1253	0.3771±0.0883	**0.0712**±0.0222	**0.1892**±0.0269	0.0713±0.0222	0.1893±0.0269	0.0713±0.0222	0.1893±0.0269
	72	0.3219±0.1336	0.4285±0.0855	**0.1237**±0.0417	**0.2548**±0.0389	0.1237±0.0418	0.2549±0.0389	0.1237±0.0418	0.2550±0.0390
	168	0.3793±0.1501	0.4697±0.0892	**0.1653**±0.0608	**0.3003**±0.0510	0.1655±0.0608	0.3005±0.0510	0.1654±0.0609	0.3006±0.0514
	336	0.4179±0.1595	0.4968±0.0909	**0.1962**±0.0689	**0.3329**±0.0538	0.1963±0.0689	0.3331±0.0538	0.1964±0.0691	0.3332±0.0544
	720	0.4951±0.1598	0.5463±0.0847	**0.2632**±0.0696	**0.3931**±0.0472	0.2634±0.0696	0.3932±0.0472	0.2635±0.0701	0.3943±0.0478
Visibility	24	1.0079±0.2200	0.6662±0.1353	0.6355±0.1253	**0.6147**±0.1007	**0.6350**±0.1261	0.6162±0.1009	0.6357±0.1255	0.6148±0.1009
	72	1.3548±0.3239	0.8147±0.1712	0.8115±0.1815	**0.7355**±0.1210	**0.8114**±0.1821	0.7366±0.1214	0.8121±0.1820	0.7357±0.1213
	168	1.5628±0.3818	0.8998±0.1939	0.9038±0.2073	**0.7948**±0.1260	**0.9037**±0.2077	0.7954±0.1262	0.9047±0.2083	0.7950±0.1265
	336	1.6512±0.4031	0.9364±0.2027	**0.9434**±0.2298	**0.8207**±0.1317	0.9447±0.2312	0.8233±0.1323	0.9446±0.2308	0.8210±0.1322
	720	1.7290±0.4237	0.9665±0.2113	**0.9801**±0.2481	**0.8417**±0.1366	0.9812±0.2488	0.8444±0.1367	0.9810±0.2485	0.8418±0.1369
Humidity	24	1.0022±0.1534	0.7505±0.0610	0.3548±0.0796	0.4378±0.0529	**0.3547**±0.0796	0.4380±0.0529	**0.3547**±0.0796	**0.4378**±0.0528
	72	1.2302±0.1741	0.8504±0.0608	**0.4939**±0.1060	**0.5356**±0.0625	0.4940±0.1060	0.5357±0.0625	**0.4939**±0.1060	**0.5356**±0.0625
	168	1.3995±0.2021	0.9172±0.0619	0.5770±0.1179	0.5909±0.0647	0.5776±0.1180	0.5918±0.0648	**0.5769**±0.1178	**0.5908**±0.0646
	336	1.5181±0.2296	0.9608±0.0656	0.6281±0.1240	0.6233±0.0663	0.6285±0.1240	0.6239±0.0664	**0.6278**±0.1239	**0.6231**±0.0662
	720	1.6328±0.2538	1.0030±0.0707	0.6723±0.1316	0.6500±0.0693	0.6726±0.1316	0.6505±0.0693	**0.6718**±0.1316	**0.6496**±0.0691

D.5 Discussion of the value of N_A in Neighbor Graph

When we use the neighbor graph G_N separately, the influence of the selection value of N_A should be discussed. In Figure 7, we respectively set N_A as 5, 10, 15, 20, and 25 to verify the variation trend of MAE and RMSE when the forecasting time step is 12. The performance will not be significantly improved when the number exceeds 10. Therefore, we consistently set N_A to 10 in the following ablation studies.

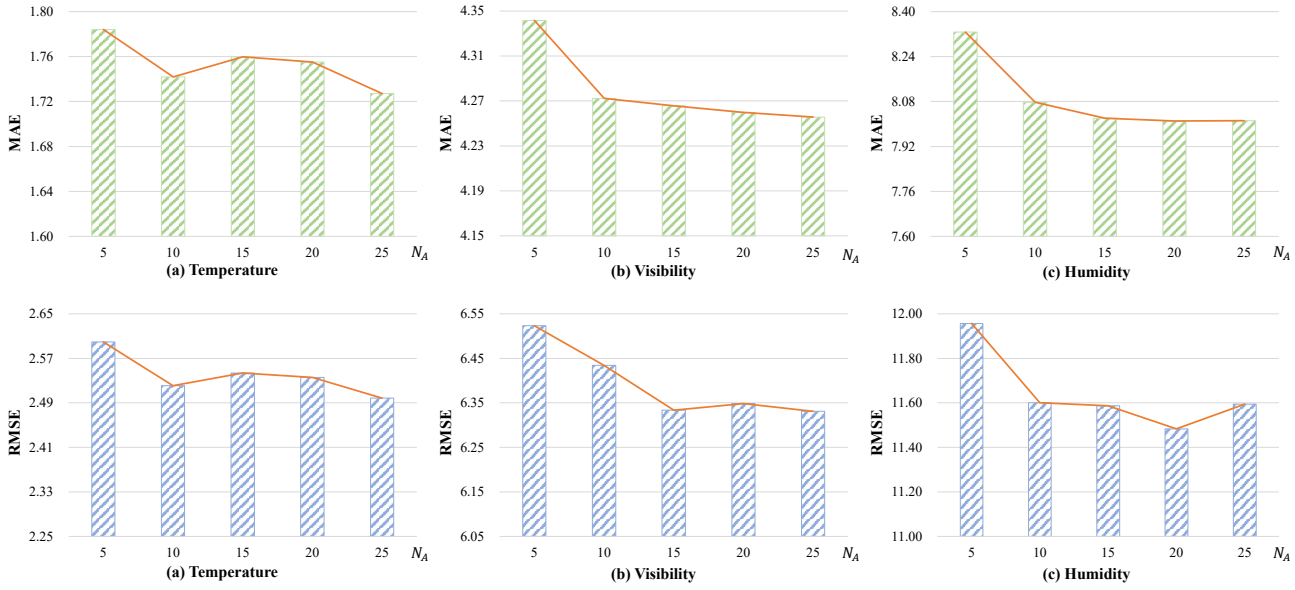

Figure 7: The selection of N_A in neighbor graph in forecasting (a) temperature, (b) visibility, and (c) humidity.