Towards Resolution-Aware Retrieval Augmented Zero-Shot Forecasting

Iman Deznabi[1,2*] **Peeyush Kumar**[2] **Madalina Fiterau**[1]
[1] University of Massachusetts Amherst [2] Microsoft Research
{iman, mfiterau}@cs.umass.edu
peeyush.kumar@microsoft.com

Abstract

Zero-shot forecasting predicts variables at locations or conditions without direct historical data, a challenge for traditional methods due to limited location-specific information. We introduce a retrieval-augmented model that leverages spatial correlations and temporal frequencies to enhance predictive accuracy in unmonitored areas. By decomposing signals into different frequencies, the model incorporates external knowledge for improved forecasts. Unlike large foundational time series models, our approach explicitly captures spatial-temporal relationships, enabling more accurate, localized predictions. Applied to microclimate forecasting, our model outperforms traditional and foundational models, offering a more robust solution for zero-shot scenarios.

1 Introduction

Zero-shot forecasting aims to predict outcomes for previously unseen locations or conditions without direct historical data. This is crucial in scenarios where acquiring location-specific data is costly or infeasible. Our approach leverages spatial correlations and temporal frequency characteristics, enabling effective zero-shot forecasting by incorporating knowledge from similar contexts.

In domains like agriculture and urban planning, precision and local accuracy are essential. Traditional forecasting models often rely on data from distant stations, leading to inaccuracies. For example, a farmer using data from a weather station 50 miles away may miss localized conditions like unexpected frost, resulting in crop damage [14]. This underscores the need for reliable localized prediction mechanisms.

We introduce a retrieval-augmented zero-shot forecasting model that enhances accuracy by utilizing the correlation between spatial proximity and temporal frequency. By decomposing environmental signals into different frequencies, our model identifies patterns in both space and time, improving the forecasting of microclimate variables. This approach is particularly valuable in fields like agriculture, ecological conservation, and urban design, where microclimates can vary significantly over short distances.

While large foundation time-series models [1, 15, 6, 8, 18] exhibit strong generalization capabilities, they struggle to adapt to new, unseen locations due to the lack of specific contextual data. Existing deep learning models for spatio-temporal weather modeling [9, 3, 19] are often computationally expensive and limited to lower spatial resolutions or global scales. Additionally, these models typically assume data points are on a grid with equal spacing, which does not hold in many microclimate scenarios. We address these issues using Graph Neural Networks (GNNs) [21, 4] to handle arbitrary distances between points. Numerical Weather Prediction (NWP) models [5, 11, 16] and adaptive learning methods [20, 17, 22] face similar limitations, lacking the capability for resolution-aware retrieval. Our model extends recent retrieval-augmented forecasting methods [10, 7] by incorporating

NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

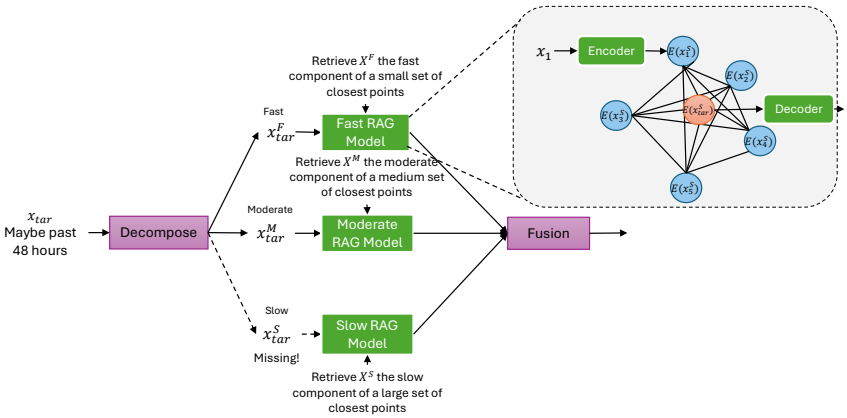

Figure 1: Model structure, we decompose the target station's preceeding past values and then use our RAG forecasting model for each frequency using more points for slower frequency signals.

resolution-aware retrieval and GNN-based adaptive learning, offering a novel approach to zero-shot forecasting.

This paper develops and evaluates the retrieval-augmented model, demonstrating its superior performance in zero-shot forecasting scenarios by effectively using both spatial and temporal information.

2 Methodology

We aim to forecast climate parameters over a time horizon L_y starting from the current time t, predicting values $\mathcal{Y}_t = \{y_{t+1}, y_{t+2}, ..., y_{t+L_y} | y_i \in \mathbb{R}\}$. We use as input a limited preceding window of climate parameters, $\mathcal{X}_t = \{x_{t-L_x}, x_{t-L_x+1}, ..., x_t\}$, with each $x_t \in \mathbb{R}^n$ representing n available climate parameters at time t. Our predictions focus on a specific location, the target station (st_{tar}), characterized by its geographic data $\ell(st_{tar})$, including latitude and longitude. The historical dataset $\mathcal{H} = \{(\mathcal{X}_{t'}, \mathcal{Y}_{t'}) | t' < t\}$ contains past climate values from multiple stations but lacks data from st_{tar} in the zero-shot scenario ($(\mathcal{X}_t(st_{tar}), \mathcal{Y}_t(st_{tar})) \notin \mathcal{H}\ \forall t$). Consequently, we rely on data from other stations and the immediate preceding window at st_{tar} to forecast at the target station. Our method addresses this zero-shot forecasting problem by developing models capable of accurate climate forecasting at st_{tar} without historical data specific to that location.

2.1 Model structure

The overall architecture of our model is illustrated in Figure 1. The core of our approach is a retrieval-augmented (RAG) forecasting model, designed to efficiently leverage both historical data and spatial information of reference points to predict values for a target point. To enhance this forecasting capability, we decompose the input signals into multiple frequency components. For each frequency, we apply a separate instance of the RAG forecasting model, utilizing a distinct set of retrieved reference points tailored to that frequency's characteristics. The outputs from these individual models are then fused to generate the final forecast.

In the following sections, we first detail the RAG forecasting model, outlining its structure and functionality. We then explain the concept of resolution-aware retrieval, providing the rationale behind selecting reference points for each frequency. Finally, the training procedure used to optimize the model is given in Appendix Section C.1.

2.2 Retrieval-Augmented Forecasting model

In the zero-shot forecasting setting, the past values (or context) available for the target point are limited. Therefore, to improve forecasting accuracy, it is crucial to leverage information from other reference points. Our model addresses this by learning the following probability distribution: $P(\mathcal{Y}_t(st_{tar}) | \mathcal{X}_t(st_{tar}), \mathcal{X}_t(st_1^{R_{tar}}), \mathcal{X}_t(st_2^{R_{tar}}), \mathcal{X}_t(st_3^{R_{tar}}), ...)$ where $\{\text{st}_1^{R_{tar}}, \text{st}_2^{R_{tar}}, \text{st}_3^{R_{tar}}, ...\} \in \psi(\text{st}_{tar})$ represents the set of reference points for the target point. The function $\psi(\text{st}_i)$ retrieves these reference points based on a distance function $d(\text{st}_i, \text{st}_j)$: $\psi(\text{st}_i) = \text{argmin}^{\text{topk}}_{\text{st}_j} d(\text{st}_i, \text{st}_j)$

For simplification, we denote the past values of the retrieved reference points as $\mathcal{X}_t^R$. The final forecast for the target point is then given by: $\hat{\mathcal{Y}}_t(st_{tar}) = \phi(\mathcal{X}_t, \mathcal{X}_t^R)$

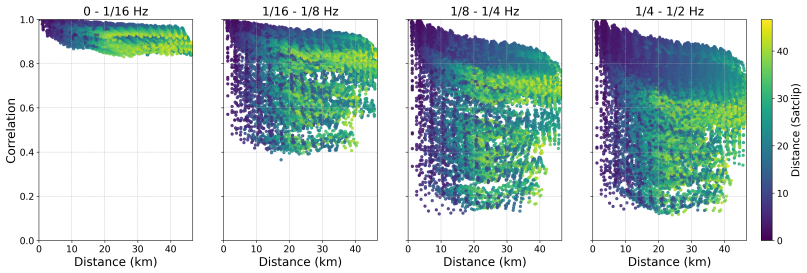

Figure 2: Correlation between historical temperature data and distance for different frequencies. The color indicates the Satclip embedding distances. In lower frequencies (left), distant points maintain high correlation, while in higher frequencies (right), correlation drops more quickly with distance.

To achieve this, our model employs a transformation module that maps the information from the neighboring reference points to the target point. After experimenting with several architectures, we found that Graph Neural Networks (GNNs) performed best. In this setup, nodes represent embeddings of the retrieved reference points, allowing for efficient information transfer and improved forecasting accuracy.

2.2.1 GNN Transform Module

Our model refines the embedding of the target location by leveraging embeddings from similar locations. We construct a graph with the retrieved reference points, where edge weights are inversely proportional to the distance between location embeddings derived from Satclip [13]: $W_{i,j} = e^{-||\mathcal{L}(st_i)-\mathcal{L}(st_j)||}$ where $\mathcal{L}(st_i)$ is the location embedding of station i. Using this graph, we employ a Graph Convolutional Neural Network (GCNN) [12] to aggregate information from neighboring locations, enhancing the target's embedding for forecasting.

2.3 Resolution-Aware Retrieval

In many spatio-temporal time series such as climate, the patterns affect surrounding areas in a way that closer locations have a greater influence on the fast-changing, higher-frequency components, while more distant locations affect slower-changing, lower-frequency components. As illustrated in Figure 2, this is evident in the correlation of points over varying distances for different frequencies. Based on this observation, we developed a model that retrieves different sets of reference points depending on the frequency of the component.

To implement this, we first decompose the context at the target location using wavelet decomposition: $\mathcal{W}(\mathcal{X}_t(st_{tar})) = \{\mathcal{X}_t^{f_1}(st_{tar}), \mathcal{X}_t^{f_2}(st_{tar}), \mathcal{X}_t^{f_3}(st_{tar}), ...\}$ where $\mathcal{W}$ represents the wavelet decomposition. The final model, incorporating this resolution-aware retrieval, is defined as: $\hat{\mathcal{Y}}_t(st_{tar}) = \mathcal{W}^{-1}(\{\phi(\mathcal{X}_t^{f_1}(\psi_{f_1}(st_{tar}))), \phi(\mathcal{X}_t^{f_2}(\psi_{f_2}(st_{tar}))), ...\})$ where $\psi_{f_i}(\text{st}_{tar})$ is the set of points retrieved for the target point st_{tar} at frequency f_i, and $\mathcal{W}^{-1}$ is the inverse wavelet transform.

We enforce the constraint: $|\psi_{f_i}(st_{tar})| > |\psi_{f_j}(st_{tar})|,\ if\ f_i < f_j$

This ensures that for lower frequencies, where the signal varies more gradually, the model retrieves a larger set of reference points. Additionally, because the wavelet decomposition captures increasingly coarser features at lower frequencies, we also have: $|\mathcal{X}_t^{f_i}| < |\mathcal{X}_t^{f_j}|,\ if\ f_i < f_j$

This design allows the model to rely on a broader context for lower frequencies, improving prediction accuracy where the available data is sparser.

3 Results

In this section, we evaluate the performance of our model using a real-world dataset. We compare our results against the baselines described in Appendix B and perform an ablation study on various model components.

We use the ERA5 dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF), focusing on 320 points in the Northwestern United States, spanning from coordinates $(44°, -120°)$ to $(49°, -124°)$. The data covers the period from 2019 to 2023. To simulate zero-shot forecasting,

Model	MSE↓	MAE↓
HRRR	11.37	2.48
Chronos base	8.33	2.11
Chronos large	8.21	2.10
TimesFM	9.21	2.21
Informer	6.69	1.91
Informer + Decomposition	6.51	1.91
Informer + Decomposition + GNN	**6.36**	**1.89**

Table 1: Results of the model on zero-shot test stations.

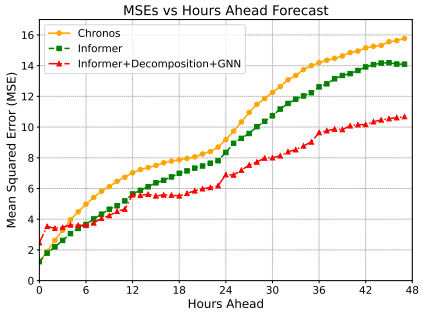

Figure 3: MSE vs. hours ahead.

we randomly selected 10 points as zero-shot locations, excluding them from the training data. The future data from these zero-shot points was used as test data.

For temporal splitting, we used the first 70% of the data from the training stations for training the models, 10% for validation (from the same training stations), and the final 20% of the data from the zero-shot points for testing. We predict 2 meter temperature as target (in degrees Kelvin). A comprehensive list of the selected features and corresponding weather stations is provided in Appendix A.

Using data from the past 96 hours ($L_x = 96$), we developed forecasts for the next 48 hours ($L_y = 48$) for the 2 meters above ground temperature. We then calculated and reported the mean squared error for each hour of the prediction.

For our current models we used Informer[23] as the encoder decoder forecaster and added our models on top of this architecture.

Table 1 presents the mean squared error (MSE) results of our model averaged over the 10 zero shot stations, along with comparisons to multiple baseline models. Notably, our model outperforms numerical weather prediction model usually used in these scenarios High Resolution Rapid Refresh (HRRR)[2], and large foundation time series models, Chronos[1], and TimesFM[6].

Additionally, we provide MSE results for different forecasting horizons in Figure 3, comparing Chronos, Informer, and our full model. Although Informer and Chronos perform better for short-term predictions (a few hours ahead), our model excels at longer forecasting horizons (10+ hours ahead). This improvement is due to our retrieval and decomposition process, which enhances the prediction of lower-frequency components, which is more critical for longer-term forecasts. In appendix section D we also show 47-hour ahead forecast comparison between HRRR, Chronos, and our model.

We also conducted an ablation study to assess the impact of different model design choices, the results of which are given in the Appendix D.1. We show how much each model component improves the performance, and the model incorporating all components performs the best.

4 Conclusion and future work

In this article, we proposed a retrieval-augmented zero-shot forecasting model that leverages reference points from well-monitored regions to enhance forecasting accuracy in unmonitored locations. Unlike large foundation time-series models [1, 15, 6, 8, 18], which often struggle to adapt to new, unseen locations due to the lack of specific contextual data, our approach incorporates a resolution-aware retrieval strategy and a GNN-based module to address this limitation. The results show that by using frequency-specific retrieval sets and advanced graph-based transformations, our model improves the forecasting performance.

For future work, we plan to explore integrating our method with other state-of-the-art encoder-decoder models like PatchTST [18] and MOMENT [8], examining if this hybrid approach can mitigate the contextual data limitations faced by foundation models. We also intend to evaluate our model on larger and more diverse datasets, including global data, to better understand its generalizability across various climates and geographical regions. Furthermore, we will conduct a comprehensive evaluation, including robustness checks and comparisons with fine-tuned large foundation models, and spatio-temporal forecasting models to rigorously assess the efficacy of our method in adapting to the unique challenges posed by zero-shot microclimate forecasting.

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

[2] Stanley G Benjamin, Stephen S Weygandt, John M Brown, Ming Hu, Curtis R Alexander, Tatiana G Smirnova, Joseph B Olson, Eric P James, David C Dowell, Georg A Grell, et al. A north american hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, 144(4):1669–1694, 2016.

[3] Alabi Bojesomo, Hasan AlMarzouqi, and Panos Liatsis. A novel transformer network with shifted window cross-attention for spatiotemporal weather forecasting. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[4] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

[5] Jean Coiffier. *Fundamentals of numerical weather prediction*. Cambridge University Press, 2011.

[6] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[7] Iman Deznabi, Peeyush Kumar, and Madalina Fiterau. Zero-shot microclimate prediction with deep learning. *arXiv preprint arXiv:2401.02665*, 2024.

[8] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

[9] Jake Grigsby, Zhe Wang, Nam Nguyen, and Yanjun Qi. Long-range transformers for dynamic spatiotemporal forecasting. *arXiv preprint arXiv:2109.12218*, 2021.

[10] Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. Retrieval based time series forecasting. *arXiv preprint arXiv:2209.13525*, 2022.

[11] Ryuji Kimura. Numerical weather prediction. *Journal of Wind Engineering and Industrial Aerodynamics*, 90(12-15):1403–1414, 2002.

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

[14] Peeyush Kumar, Ranveer Chandra, Chetan Bansal, Shivkumar Kalyanaraman, Tanuja Ganu, and Michael Grant. Micro-climate prediction-multi scale encoder-decoder based deep learning framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3128–3138, 2021.

[15] Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6555–6565, 2024.

[16] Andrew C Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

[17] Huiming Lu, Jiazheng Wu, Yingjun Ruan, Fanyue Qian, Hua Meng, Yuan Gao, and Tingting Xu. A multi-source transfer learning model based on lstm and domain adaptation for building energy prediction. *International Journal of Electrical Power & Energy Systems*, 149:109024, 2023.

[18] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[19] 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.

[20] Mohamed Ragab, Emadeldeen Eldele, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Self-supervised autoregressive domain adaptation for time series data. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1341–1351, 2022.

[21] 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, 2020.

[22] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[23] 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.

Hyperparameter	Value
batchsize	32
inner model embedding size	2048
embedding size	256
dropout	0.05
learning rate	0.0001
loss	mse
number of heads	8
patience	10
encoder layers	2
decoder layers	1
Biggest Graph Size	15
Graph Size Reduction	5
GCNN conv layers	2
GCNN hidden dim	256
Wavelet decomposition levels	6

Table 2: Hyperparameters for our model

A ERA5 data

We downloaded these 5 features:

- **u10**: The eastward component of the wind at 10 meters above the ground. It represents the wind speed in the east-west direction.
- **v10**: The northward component of the wind at 10 meters above the ground. It represents the wind speed in the north-south direction.
- **t2m**: Temperature at 2 meters above the ground.
- **d2m**: Dew point temperature at 2 meters above the ground.
- **sp**: Surface pressure, which is the atmospheric pressure at the Earth's surface.

hourly from January 1, 2019, to December 31, 2023 for gridded points between coordinates $(44°, -120°)$ and $(49°, -124°)$ at $0.25°$ intervals.

B Contenders

We compare our models against these contending forecasting models:
Chronos[1]: We used the pre-trained base with 200m parameters and large with 710m parameters to generate forecasts for our data. We also tried to fine-tune the Chronos base model on our data but could not improve the results.
TimesFM[6]: We tried the shared pre-trained model with 200m parameters also as suggested by authors we tried giving different date and time based parameters as covariates as well as forecasting other features using the same model and feeding them as dynamic numerical covariates, however, the best performing model in our case was without using any covariates and the single variate original model.
HRRR[2]: The High-Resolution Rapid Refresh (HRRR) model is a numerical weather prediction model that provides high-resolution, frequently updated forecasts for the contiguous United States, using real-time weather data to deliver detailed predictions on an hourly basis. We downloaded the forecasts for all the locations from National Oceanic & Atmospheric Administration (NOAA) website.

C Model hyperparameters

The hyperparameters used for ERA5 dataset for Informer and our model are given in Table 2, we select these hyperparameters using 10% of data that come after the training set and before the test set.

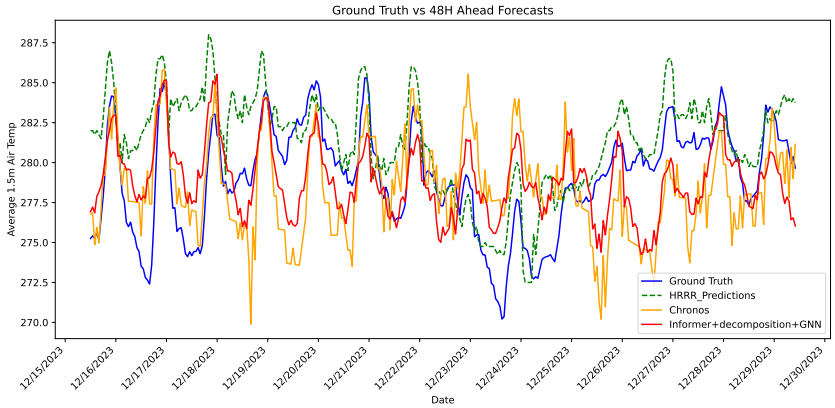

Figure 4: Predictions of our best zero-shot model compared with HRRR predictions and ground truth on the last two weeks of December 2023.

C.1 Training procedure

Our training methodology consists of two distinct phases. In the first phase, we train the encoder-decoder model to forecast the data using all available training stations. During this phase, the model learns to capture the global patterns and relationships across the entire dataset, and the parameters are updated accordingly.

In the second phase, we refine the weights of the GCNN module to enhance the performance of the RAG forecasting model. We designate each training station as the target station, freezing all previously learned model parameters except for those in the GNN module. By training only the GNN weights, the model learns to efficiently transfer information from neighboring stations, enabling more accurate forecasting for the target station.

D Other Results

In Figure 4 we show the 48 hours ahead forecasts of HRRR, Chronos and our model vs the ground truth for the last two weeks of December 2023.

D.1 Ablation study

In Table 3 we provide the forecasting performance of our model with different parts removed on the validation data of a random zero-shot station. We show results with averaging the forecasts of 5 closest stations (Informer + Average of 5 forecasts), using a Transformer for RAG forecast model instead of a GNN (Informer + Transformer), using GNN model (Informer + GNN), using Informer with wavelet decomposition (Informer + Decomposition), using decomposition and GNN with the same size (Informer + Decomposition + GNN with same graph sizes) and finally our full model which uses decomposition and GNN with different graph sizes for different frequencies. You can see that our full model achieves the best performance with all the parts included.

Model	MSE↓	MAE↓
Informer	8.03	2.22
Informer + Average of 5 forecasts	7.66	2.12
Informer + Transformer	7.26	2.07
Informer + GNN	7.19	2.06
Informer + Decomposition	7.17	2.10
Informer + Decomposition + GNN with same graph sizes(5)	7.44	2.09
Informer + Decomposition + GNN with same graph sizes(15)	7.18	2.08
Full model (Informer + Decomposition + GNN with different graph sizes)	6.89	2.04

Table 3: Ablation study on data of one random location