AirCast: Improving Air Pollution Forecasting Through Multi-Variable Data Alignment

Vishal Nedungadi[1,2], Muhammad Akhtar Munir[1], Marc Rußwurm[2], Ron Sarafian[3], Ioannis N. Athanasiadis[2], Yinon Rudich[3], Fahad Shahbaz Khan[1,4], Salman Khan[1,5]

[1] **Mohamed bin Zayed University of AI, UAE**
[2] **Wageningen University and Research, The Netherlands**
[3] **Weizmann Institute of Science, Israel**
[4] **Linköping University, Sweden**
[5] **Australian National University, Australia**

Abstract

Air pollution remains a leading global health risk, exacerbated by rapid industrialization and urbanization, contributing significantly to morbidity and mortality rates. In this paper, we introduce *AirCast*, a novel multi-variable air pollution forecasting model, by combining weather and air quality variables. *AirCast* employs a multi-task head architecture that simultaneously forecasts atmospheric conditions and pollutant concentrations, improving its understanding of how weather patterns affect air quality. Predicting extreme pollution events is challenging due to their rare occurrence in historic data, resulting in a heavy-tailed distribution of pollution levels. To address this, we propose a novel Frequency-weighted Mean Absolute Error (fMAE) loss, adapted from the class-balanced loss for regression tasks. Informed from domain knowledge, we investigate the selection of key variables known to influence pollution levels. Additionally, we align existing weather and chemical datasets across spatial and temporal dimensions. *AirCast*'s integrated approach, combining multi-task learning, frequency weighted loss and domain informed variable selection, enables more accurate pollution forecasts.

1. Introduction

Rapid industrialization, economic growth, and climate change have significantly worsened air pollution (Nakhjiri and Kakroodi, 2024), raising serious concerns about environmental quality and public health. Among various pollutants, particulate matter such as PM1, PM2.5, and PM10 (particles smaller than 1, 2.5, and 10 micrometers, respectively) have been directly associated with adverse health effects. These tiny particles can penetrate the respiratory system, possibly leading to cancer and various respiratory and cardiovascular diseases. The World Health Organization (WHO) reports that around 99% of the global population is exposed to air that does not meet its 2019 quality guidelines. According to recent estimates (WHO, 2024), air pollution is responsible for approximately 6.7 million premature deaths annually. This highlights the urgent need for improved forecasting methods to accurately predict air quality. These forecasts can advise policy decisions and contribute to reducing emissions strategies.

Air pollution forecasting for PM primarily relies on two approaches: physics-based and data-driven models. Physics-based models

© 2025 .

simulate pollutant dispersion and chemical transformations using fundamental principles of atmospheric chemistry and physics. These models often use non-linear empirical methods (Cobourn, 2010; Lv et al., 2016) to represent complex environmental interactions. Although they offer valuable insight into the physical and chemical processes that govern air quality, their accuracy is often constrained by the dynamic complexity of atmospheric systems. This makes it difficult to precisely capture both long and short-term trends. In contrast, data-driven approaches (Bi et al., 2023; Nguyen et al., 2023a,b; Bodnar et al., 2024) utilize machine learning methodologies to model complex relationships among various atmospheric variables, such as temperature, wind, and PM concentrations. These models are trained to capture non-linear patterns and dependencies implicitly under diverse atmospheric conditions. Moreover, data-driven models adapt more readily to new data and evolving environmental conditions than physics-based models, identifying patterns and relationships that physics-based models cannot explicitly represent. Existing data-driven approaches overlook variables that could potentially influence PM concentrations.

In this work, we enhance PM forecasting methods by integrating weather and air quality variables to improve accuracy. Our proposed model, AirCast, is a Vision Transformer (ViT) (Dosovitskiy et al., 2021), designed for air pollution forecasting by adapting a weather foundational model Nguyen et al. (2023a). By utilizing large-scale pretrained models, AirCast learns generalizable representations from diverse datasets, enhancing its performance in air quality prediction. An important aspect of this adaptation is our development of a combined dataset integrating weather and air quality variables for precise air pollution forecasting. We source weather variables from WeatherBench (Rasp et al., 2020) and air quality variables from the Copernicus Atmosphere Monitoring Service (CAMS) EAC4 dataset (ECMWF, 2023). This multi-variable approach allows AirCast to capture the complex relationships between weather conditions and pollutant levels. Similar to Nguyen et al. (2023a), the model architecture incorporates variable tokenization and variable aggregation modules to efficiently handle a large number of variables and reduce the sequence length. To further enhance its capabilities, a multi-task head architecture enables the model to predict both atmospheric weather and air pollution variables simultaneously. Additionally, a Frequency-weighted Mean Absolute Error (fMAE) loss function inspired by the class balanced loss function (Cui et al., 2019) addresses the heavy-tailed distributions of pollutants, improving the accuracy of predictions for extreme cases. Furthermore, learning from domain knowledge we also investigate the selection of key variables known to affect PM concentrations.

Our study focuses on the Middle East and North Africa (MENA) region, which consistently experiences some of the highest levels of PM concentrations globally, often exceeding the recommended air quality standards of the WHO (Heger et al., 2022). Forecasting air pollution in the MENA region using data-driven methods is extremely important due to its distinct environmental challenges. The challenges include frequent dust storms, industrial emissions, reliance on fossil fuels, rapid urbanization, and low rainfall, all of which combine to significantly degrade air quality. In this paper, we focus on accurate and efficient PM forecasting in MENA region, aiming to support mitigation efforts and reduce the harmful effects of air pollution. Our main contributions are as follows:

1. **Integrated Forecasting:**: To capture the interactions between weather and air quality, we develop a multi-task head

Vishal Nedungadi, M. Akhtar Munir, M. Rußwurm, R. Sarafian, I. Athanasiadis, Y. Rudich, F. S. Khan, S. Khan

architecture that simultaneously predicts atmospheric and pollution variables.

2. **Frequency-weighted Loss Function:** To address the heavy-tailed distributions of pollutants like PM1, PM2.5, and PM10, we introduce a Frequency-weighted Mean Absolute Error (fMAE).

3. **Regional Adaptation:** Recognizing the MENA region's high PM concentrations, we enhance the model to improve the accuracy of forecasts for severe pollution levels in the region.

4. **Combined Dataset:** To help the model learn the relationships between atmospheric conditions and pollutant levels, we create a comprehensive dataset by aligning existing weather and chemical datasets across spatial and temporal dimensions.

2. Related Work

In recent years, the integration of machine learning methods into various scientific domains has gained significant attention, with air pollution forecasting a notable example. Traditionally, physics-based models like the WRF-Chem (Ojha et al., 2020) and CMAQ (Zhang et al., 2012) model have been employed to predict air pollution levels. These models are grounded in the fundamental principles of atmospheric chemistry and physics to simulate complex interactions within the atmosphere. However, the highly nonlinear and complex nature of air pollution variables poses substantial challenges for these physics-based models. Modeling air pollution's complexity often leads to high uncertainties and reduced prediction accuracy (Hao et al., 2020; Li et al., 2019). Additionally, running these models at high resolutions in complex environments demands significant computational resources, which can be challenging for real-time forecasting.

In contrast, data-driven machine learning models (Yu et al., 2022; Cai et al., 2023; Bodnar et al., 2024) have emerged as a more effective alternative for air pollution forecasting. These models excel at capturing nonlinear relationships and patterns within large datasets, allowing them to handle the complexities of air pollution variables more adeptly. Machine learning approaches can provide more accurate and efficient predictions without requiring detailed physical simulations by learning directly from the data. In Yu et al. (2022), a deep ensemble-based approach is introduced for estimating daily PM2.5 concentrations. This framework leverages machine learning base models, such as XGBoost, which are used to train meta-models in the second stage, with an optimization algorithm applied in the third stage. In Cai et al. (2023), authors proposed a framework to enhance the prediction of hourly PM2.5 concentrations. Their method involves breaking down complex data into simpler components, each representing different frequency levels. These components are then modeled using a combination of autoregressive and CNN-based methods to capture patterns in data, to improve the accuracy of the predictions.

Recent advancements in neural network architectures have significantly improved the integration of air quality variables with weather variables in forecasting models. For instance, the Aurora model (Bodnar et al., 2024) primarily trains on weather data and then forecasts air pollution levels as a downstream task. Similarly, ClimaX (Nguyen et al., 2023a) is an open-source weather model that leverages the vision transformer architecture (Dosovitskiy et al., 2021). Trained on large-scale datasets, it serves as a foundational model by employing a pretext task focused on predicting future time steps randomly sampled within a specified range. In its downstream applications, ClimaX handles a variety of tasks across different spatial and temporal

Table 1: A list of all the weather and air quality variables present in our dataset. Furthermore, for variables that contain data at different pressure levels, we collect 7 of them.

	Variable (short name)	Description	Pressure Levels
Weather Variables	geopotential (z)	Varies with the height of a pressure level	7 levels
	temperature (t)	Temperature	7 levels
	specific humidity (q)	Mixing ratio of water vapor	7 levels
	relative humidity (r)	Humidity relative to saturation	7 levels
	u component of wind (u)	Wind in longitude direction	7 levels
	v component of wind (v)	Wind in latitude direction	7 levels
	2m temperature (t2m)	Temperature at 2m height above surface	Single level
	10m u component of wind (u10)	Wind in longitude direction at 10m height	Single level
	10m v component of wind (v10)	Wind in latitude direction at 10m height	Single level
Air Quality Variables	carbon monoxide (co)	Carbon monoxide concentrations	7 levels
	ozone (go3)	Ozone concentrations	7 levels
	Nitrogen monoxide (no)	Nitrogen monoxide concentrations	7 levels
	Nitrogen dioxide (no2)	Nitrogen dioxide concentrations	7 levels
	Sulphur dioxide (so2)	Sulphur dioxide concentrations	7 levels
	Particulate matter d <1 µm (pm1)	Particulate matter with diameter less than 1 µm	Single level
	Particulate matter d <10 µm (pm10)	Particulate matter with diameter less than 10 µm	Single level
	Particulate matter d <2.5 µm (pm2.5)	Particulate matter with diameter less than 2.5 µm	Single level
	Total column carbon monoxide (tcco)	Total amount overall levels	Single level
	Total column nitrogen monoxide (tc_no)	Total amount overall levels	Single level
	Total column nitrogen dioxide (tcno2)	Total amount overall levels	Single level
	Total column ozone (gtco3)	Total amount overall levels	Single level

scales, including regional weather forecasting. Due to its adaptable and efficient design, we have selected ClimaX to demonstrate our proposed approach. A recent study (Munir et al., 2024), has explored enhancing ClimaX for MENA weather forecasting using parameter-efficient fine-tuning like LoRA. This underscores the potential of adapting foundational models to meet the challenges of specific regions. However, it's important to note that ClimaX currently operates at a lower spatial resolution compared to models like Aurora (Bodnar et al., 2024) and CAMS (ECMWF, 2023). In contrast to large-scale foundational models, existing work on PM forecasting typically employs fewer variables, limiting itself to smaller capacity models or statistical approaches (Cabello-Torres et al., 2022; Masood et al., 2023). A notable exception is the work of Sarafian et al. (2023), which uses a transformer-based model to forecast PM10 concentrations using several weather variables, demonstrating their importance in the process. However, most studies in this field, while valuable, often focus on predicting a single PM concentration variable and utilize only a subset of available predictors. Our work aims to address these limitations by adopting a more comprehensive, multi-variable approach.

In AirCast, we combine air quality and weather data to better capture complex environmental dynamics, addressing the limitations of models that rely on one type of data. This approach is important for regions like the MENA, where diverse environmental factors require models capable of handling intricate interactions effectively.

3. Aligned Dataset

We aim to create an aligned dataset with multiple weather and air quality variables (see Table 1) to better capture the complex factors influencing PM concentrations.

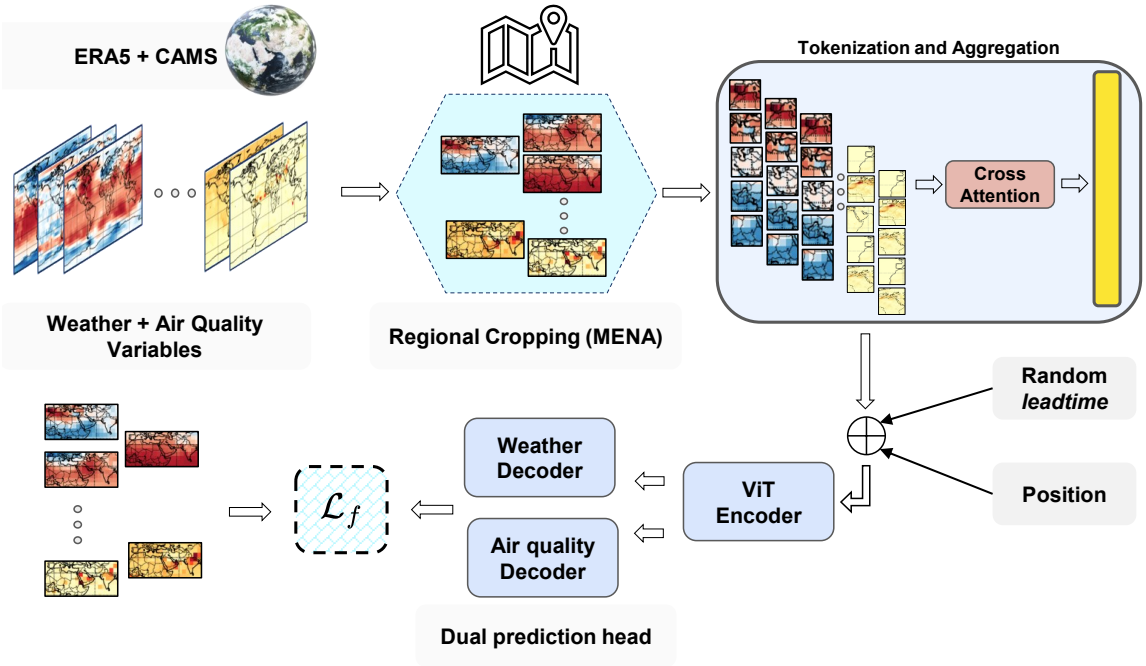


Figure 1: This illustrates the architecture of the **AirCast** model, an extension of Nguyen et al. (2023a). The model integrates weather data from the ERA5 dataset and air quality data from the CAMS EAC4 dataset. The model is trained using regional data from the MENA region. The input variables are tokenized and aggregated, with a Vision Transformer (ViT) encoder, processing the combined weather and air quality inputs. A dual decoder head is employed, with one predicting weather variables and the other forecasting air quality variables. The predictions are compared with the ground truth at a certain lead time using the Frequency-Weighted MAE loss function.

Weather Variables. The weather data is sourced from the ERA5 archive (Rasp et al., 2020), providing hourly data from 1979 to 2018. Due to its large size, the dataset has been regridded to resolutions of 5.625^o (32×64 pixels), and 1.40525^o (128×256 pixels). Furthermore, to temporally align with the chemical pollutant variables described next, we only choose the years from 2003 to 2018. For our experiments, we focus on the 5.625^o resolution to balance data granularity and computational efficiency.

Air Quality Variables. The air quality data is collected from the Copernicus Atmosphere Monitor Service (CAMS) data archive. We utilize the ECMWF Atmospheric Composition Reanalysis 4 (EAC4) data catalog, which combines an atmospheric model with real-world observations to create a comprehensive global dataset comprising various air quality variables. The data originally comes in a 0.75^o resolution and three-hourly intervals. For consistency with weather data (Rasp et al., 2020), we regrid these to 5.625^o resolution. Furthermore, to align the air quality variables temporally with the weather variables, we interpolate the data to be hourly instead of three hours. Following WHO guidelines (WHO, 2021), we included additional variables known to affect PM concentrations. The full list of air quality variables is shown in Table 1.

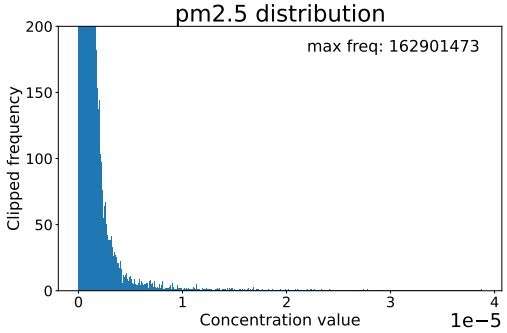


Figure 2: **Skewed distribution of PM2.5.** The y-axis corresponds to the frequency clipped at 200 (the maximum frequency is shown in each figure).

The combined dataset contains both surface and pressure-level variables. For pressure-level data, we selected seven pressure levels: 50, 250, 500, 600, 700, 850, and 925 hecto-Pascals (hPa). These levels were chosen to represent a broad range of atmospheric dynamics, from near-surface to higher altitudes. The unit hPa is typically used to represent different vertical levels in the atmosphere, with a pressure of approximately 1000 hPa at sea level, decreasing as altitude increases.

Distribution Skew. Many air quality variables show heavy-tailed distributions, notably for PM concentrations, as shown in Figure 2. This phenomenon indicates that while high pollution levels, including PM1, PM2.5, and PM10, are rare, they have a significant impact when they occur. These elevated concentrations, though infrequent, are critical indicators of severe air quality issues.

4. AirCast

In this section, we describe our approach AirCast (Figure 1), for multi-variable air pollution forecasting. Inspired by ClimaX (Nguyen et al., 2023a), we use a Vision Transformer ViT (Dosovitskiy et al., 2021) as the backbone. The model was pre-trained using a variety of climate and weather datasets, each with a varying number of variables Nguyen et al. (2023a). Similar to ClimaX, the variable tokenization module was utilized to standardize the input and a variable aggregation module was employed to handle the large sequence of input variables during training, thereby reducing the sequence length and enhancing computational efficiency.

Variable Tokenization is a process that converts each input variable separately into a sequence of patches. Specifically, each input variable V of size $H \times W$ is tokenized in a sequence of size $H/p \times W/p$, where p denotes the size of the patch. The input patches are then passed through an embedding layer, resulting in a sequence of dimensions $V \times H/p \times W/p \times D$, where D denotes the embedding dimension.

Variable Aggregation follows the variable tokenization, using a cross-attention mechanism to aggregate information from multiple variables at the same spatial location. This process effectively reduces the sequence length to $(H/p \times W/p)$ while retaining essential information from all input variables. This aggregation not only optimizes computational efficiency but also improves the model's ability to understand the relations between weather and air quality variables.

Since we aim to use weather and air quality variables, an additional prediction head is added. Both prediction heads output the same number of variables as the input. The loss is calculated independently for each set of variables: weather loss is computed for the weather variables from the weather head, and likewise for the air quality head. By decoupling the learning processes, the model alleviates potential negative transfer between tasks, which is usually a challenge in multi-task learning frameworks. Experimental evidence shows that this configuration yields the best performance.

4.1. Regional Setup

We target a specific region instead of forecasting the entire globe. While our model can be adapted globally, we focus on the MENA region to evaluate its forecasting capabilities due to the region's high PM concentrations. The MENA region consistently records some of the highest PM levels worldwide (Li et al., 2022; Nissenbaum et al., 2023), frequently exceeding WHO guidelines. We expect that by choosing such a region, we can focus our model capability on forecasting the higher PM concentrations.

4.2. Normalization

Prior to the model training, weather variables are normalized, while air quality variables undergo normalization followed by a scaled log transformation (shown in equation 1). The log transformation highlights smaller values often overshadowed by larger ones, stabilizing training and capturing the variability of low air quality concentrations more effectively. This log transformation is inverted during validation and test time.

$$x = \frac{\log(\max(x, 10^{-4})) - \log(10^{-4})}{\log(10^{-4})} \tag{1}$$

4.3. Randomized Lead Time

While our experiments focus on forecasting the variables, 24 hours from the input time (lead time), we find that randomizing the lead time during training results in improving the model performance. We believe this acts as an extra augmentation technique that may serve as a regularization, by exposing the model to various forecasting horizons. For each training sample, the lead time is randomly chosen from 6, 12, and 24 hours intervals. For validation and testing, however, only a 24-hour lead time is used to maintain consistency.

4.4. Frequency-Weighted Mean Absolute Error

Many air quality variables, including PM1, PM2.5, and PM10, exhibit a heavy-tailed distribution (as illustrated in Figure 2 for PM2.5). To address this skewness, we propose a Frequency-weighted Mean Absolute Error (fMAE) function motivated by class-balancing approaches (Cui et al., 2019). The frequency of values for each air quality variable in the training data is pre-computed using optimal bin widths specified by the Freedman-Diaconis Estimator (Freedman and Diaconis, 1981). Based on this frequency, a weight is assigned according to Equation 2,

$$W_{freq} = \begin{cases} 0\ ,\ freq = 0 \\ \frac{1-\beta}{1-\beta^{freq}},\ otherwise \end{cases} \tag{2}$$

where β is a hyperparameter that is used to define the frequency weighting term. $\beta \to 0$ signifies equal weighting while $\beta \to 1$ signifies inverse frequency weighting. Based on experimentation, we found that setting β to 0.8 resulted in the best performance. The core idea behind this weighting scheme is to provide greater emphasis on rare events and reduce the impact of frequently occurring events.

Additionally, following the methodology used in WeatherBench (Rasp et al., 2020), we further employ the latitude weight along with the frequency-weighted loss. The latitude weight is defined in Equation 3 and is used to account for the varying sizes of grid cells due to the Earth's spherical shape.

$$W_{lat}^{i} = \frac{\cos(lat(i))}{\frac{1}{H}\sum_{i`=1}^{H}\cos(lat(i`))}, \tag{3}$$

Where $lat(i)$ is the latitude of the ith row in the grid, H corresponds to the height of the image and W_{lat} is the latitude weight for each i. The overall loss is described in Equation 4

$$\mathcal{L}_f = (W_{lat} \times MAE_{weather}) + (W_{freq} \times W_{lat} \times MAE_{chemical}) \quad (4)$$

This dual-weighting approach helps the model sufficiently capture the weather and air quality variables' spatial and distributional variations. By including latitude weights, the model accounts for the variations in grid cell areas at different latitudes, which is crucial for global-scale modeling. The frequency weights, on the other hand, address the imbalance in the distribution of air quality concentrations, enhancing the model's ability to predict rare events.

With these improvements, AirCast provides an adaptive framework for multi-variable air pollution forecasting, leveraging weather and air quality data to improve accuracy in high-pollution regions.

5. Experimental Setup

5.1. Implementation details

For training AirCast on the new combined dataset, the network is initialized with pre-trained weights from Nguyen et al. (2023a). We use a learning rate of 5×10^{-4}, a batch size of 32 and a seed of 42 when training the model. The original shape of the input is 32×64 at 5.625^o resolution and after applying the regional cropping, the input is 8×14. The model is trained for 100 epochs with early stopping criteria to prevent overfitting. The training is conducted on four A100 GPUs, taking approximately four hours for the model with all variables. The dataset was temporally partitioned to create train, validation, and test splits. Specifically, data from 2003 to 2015 was allocated to the training set, while 2016 data constituted the validation set. The test set comprised data from 2017 and 2018.

5.2. Baselines

To evaluate AirCast's performance, we compare it against two established models: a persistence baseline and the CAMS global atmospheric composition forecast. Notably, to the best of our knowledge, only one other study (Aurora, Bodnar et al. (2024)) has attempted to forecast all three PM variables using a single model. However, their repository does not provide access to fine-tuned models for air pollution forecasting, precluding a direct comparison.

Persistence Baseline: This baseline model predicts that the forecast over the next 24 hours will remain unchanged from the current input. While simple, it is a valuable benchmark for evaluating the performance of more complex forecasting models.

CAMS Global Forecasts: The CAMS global atmospheric composition forecast is a comprehensive data catalog that provides twice-daily forecasts for various lead times. The forecast is generated by using a physics based atmospheric model that learns the complex patterns of several concentrations.

We conduct our baseline evaluations exclusively on data from 2017 and 2018. We standardize the input time to 00:00 and use a 24-hour lead time for all forecasts.

6. Results

In this section, we analyze the results of the air pollution forecasting experiments, focusing on PM1, PM2.5, and PM10 concentrations. We systematically examine the effects of various input variables, data transformations, and model configurations on forecasting performance. We provide the RMSE and Anomaly Correlation Coefficient (ACC) metrics (see Appendix) for most of our key ablations and just use RMSE for the others.

Impact of fMAE Loss on PM Forecasting: Considering the heavy-tailed distribution of several air quality variables, an

Table 2: AirCast Ablations: Various ablations that result in the best performing setting. The reported metrics are Root Mean Squared Error (RMSE, lower is better) and Anomaly Correlation Coefficient (ACC, higher is better). The RMSE unit for all PM concentration variables is μgm^{-3}. (*3 PM* correspond to PM2.5, PM10, PM1. The ablations in all tables use a lead time of *24 hrs* during testing. Air Quality (*AQ*) corresponds to the full list of air quality variables as shown in Table 1, which includes the 3 PM variables. *Surface (S.)* corresponds to the near-surface pressure level of multi-level variables (high pressure). ¬ is used when we consider the low pressure levels of multi-level variable. For each table, the initial setting (to compare against) is defined in gray. The best setting is defined in yellow. Abbreviations: W = Weather, S.W. = Surface Weather, AQ = Air Quality, S.AQ = Surface Air Quality.

Impact of the fMAE loss

Method	PM2.5		PM10		PM1	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
w/o fMAE	10.05	76.4	15.34	73.8	7.38	81.1
with fMAE	**9.63**	**78.2**	**14.78**	**75.5**	**7.17**	**82.4**

Adding additional variables

Variables	PM2.5		PM10		PM1	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
3 PM	9.63	78.2	14.78	75.5	7.17	82.4
W + 3 PM	9.94	77.1	14.72	76.3	7.69	79.7
AQ	9.80	78.1	14.96	75.8	7.41	81.7
W + AQ	**9.45**	**79.8**	**14.15**	**78.6**	7.24	**82.8**

Considering near surface variables.

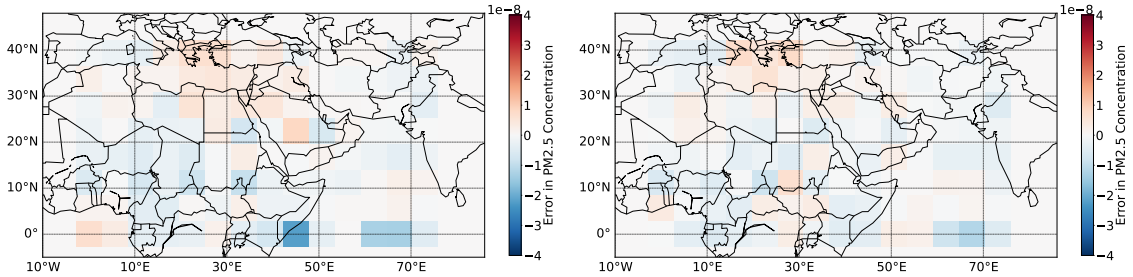
Variables	PM2.5		PM10		PM1	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
W + AQ	9.45	79.8	14.15	78.6	7.24	82.8
S.W. + AQ	9.24	80.5	13.81	79.3	7.14	82.7
W + S.AQ	9.65	79.7	14.08	79.4	7.60	82.0
S.W. + S.AQ	**8.82**	**82.7**	**13.27**	**79.7**	**6.65**	**85.3**
¬S.W. + ¬S.AQ	9.40	79.7	14.09	78.4	7.19	82.3

Baseline comparison

Method	PM2.5		PM10		PM1	
	RMSE	ACC	RMSE	ACC	RMSE	ACC
Persistence	12.08	69.0	18.86	64.4	8.61	76.3
CAMS	20.33	60.0	31.07	55.0	16.21	65.0
S.W. + S.AQ	8.82	**82.7**	13.27	**79.7**	6.65	**85.3**

experiment was conducted with and without the proposed fMAE loss. Results from Table **??** indicate that using fMAE loss led to an improvement of forecasting RMSE of 4.18%, 3.65% and 2.85% for PM2.5, PM10, PM1 respectively. For this particular experiment, we report numbers using only the three PM concentration variables as both input and output. Furthermore, visualizations using the fMAE loss (Figure 3) indicate a slightly better forecasting model for higher PM concentrations, as denoted by the lighter blue areas in the second plot.

Impact of Weather and Air Quality Inputs on PM Forecasting: Table **??** presents results from an experiment analyzing the impact of incorporating additional weather and air quality variables. Notably, air quality variables (in Table 1) include these PM concentrations as well, and we denote them as AQ. Incorporating all variables improved the forecasting RMSE by 1.87%



(a) Without fMAE loss (b) With fMAE loss

Figure 3: Sample error plots for PM2.5 forecasting (prediction - ground truth). The unit is $kg\,m^{-3}$. The first and second plots are without and with the proposed fMAE loss function respectively.

for PM2.5 and 4.26% for PM10 but slightly degraded performance for PM1. While an improvement in PM1 forecasting was anticipated, the results for PM2.5 and PM10 align with existing literature, highlighting the strong correlation between weather variables and PM concentrations (Cabello-Torres et al., 2022; Yang et al., 2017). Additionally, WHO guidelines (WHO, 2021) underscore the links between various air quality variables and PM concentrations.

Effect of Selecting Near-Surface Variables: As shown by Li et al. (2017); Sarafian et al. (2023), near-surface-level variables are crucial for forecasting PM concentrations. To verify this, we conduct experiments (Table **??**) using only near-surface-level variables from multi-level data while directly including all single-level variables. The resulting forecasting RMSE is improved by 6.67%, 6.22%, and 8.15% for PM2.5, PM10, PM1 respectively, when considering surface-level weather and air quality variables. We conduct another experiment where we consider the low-pressure level variables (represented by ¬). The results confirm that selecting variables strongly correlated with PM concentrations significantly enhances forecasting accuracy.

Baselines Comparison with Our Best Model: As mentioned earlier, no single model currently forecasts all PM concentration variables, and a direct comparison with Aurora is also not possible due to the unavailability of its fine-tuned model for air quality forecasting. Therefore, we use the persistence baseline, a simple yet effective benchmark, and CAMS global forecasts for comparison (see Table **??**).

Extreme Event Forecasting: To test our models' capability in forecasting extreme events, we selected two dust storm events: one in Kuwait on October 31, 2017, and another in Saudi Arabia on October 29, 2017. While there is room for improvement, our model demonstrated the ability to detect these events. This can be observed in Figure 4, where light blue or white colors in the affected regions closely correspond to the actual dust storm occurrences.

Randomized lead time: Following Nguyen et al. (2023b), we investigate the benefit of randomizing the lead time during training and validation. Results from table 3 indicate that randomizing lead time improves the PM forecasting performance. We believe this acts as an extra augmentation technique and allows the model to learn from various forecasting horizons.

Varying lead times: To test the models forecasting performance at various temporal times, we run an additional experiment by varying the lead time. Results from Table

Table 3: **Randomized lead time.** An ablation to test the performance with and without the randomized lead time (6, 12, 24 hrs) during training and validation. At test time, the lead time is fixed to 24hrs.

Method	RMSE (μgm^{-3}) PM2.5 / PM10 / PM1
Without	9.04 / 13.61 / 6.80
With	**8.82** / **13.27** / **6.65**

Table 4: **Varying lead time.** Considering only near surface weather and air quality variables.

Lead time	RMSE (μgm^{-3}) PM2.5 / PM10 / PM1
6 hrs	6.30 / 9.26 / 5.04
12 hrs	7.60 / 11.2 / 6.03
24 hrs	8.82 / 13.27 / 6.65
48 hrs	11.50 / 18.01 / 7.97

4 indicate that there is an inverse relation between the lead time and forecasting RMSE. This suggests that our models forecasting ability is more robust in near-term.

7. Conclusion

Previous efforts in air pollution forecasting have primarily relied on statistical models, traditional machine learning approaches, or limited variable sets. In this work, we propose a multi-variable approach with a particular focus on forecasting PM concentrations. We develop a spatially and temporally aligned dataset that integrates chemical pollutant and weather data. Building on this, we introduce AirCast, a Vision Transformer (ViT)-based forecasting model that leverages these diverse variables. Our results demonstrate that incorporating weather and air quality variables significantly enhances PM forecasting accuracy. Notably, near surface-level variables emerge as the most impactful in driving the synergy between weather and air quality data. To address the heavy-tailed distribution of chemical variables, we introduce a Frequency-weighted Mean Absolute Error (fMAE) loss function.

Impact Statement

Accurate air pollution forecasting is crucial for protecting public health and informing environmental policy decisions. From a public health perspective, reliable forecasts enable individuals with respiratory conditions or other sensitivities to take precautionary measures, reducing their exposure to harmful pollutants. This approach can potentially lead to decreased healthcare costs and an improved quality of life for vulnerable populations. Furthermore, precise forecasting empowers government and healthcare systems to better prepare for and respond to air pollution events. By anticipating periods of poor air quality, authorities can implement timely interventions, such as issuing public health advisories or temporarily restricting high-emission activities. While the method described in this paper is only for a relatively short lead time, this sets the road for future work that can improve forecasts for longer periods of time.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970): 533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024.

Rita Jaqueline Cabello-Torres, Manuel Angel Ponce Estela, Odón Sánchez-Ccoyllo, Edison Alessandro Romero-Cabello, Fausto Fernando García Ávila, Carlos Alberto Castañeda-Olivera, Lorgio Valdiviezo-Gonzales, Carlos Enrique Quispe Eulogio, Alex Rubén Huamán De La Cruz, and Javier Linkolk López-Gonzales. Statistical modeling approach for pm10 prediction before and during confinement by covid-19 in south lima, perú. *Scientific Reports*, 12(1), October 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20904-2. URL http://dx.doi.org/10.1038/s41598-022-20904-2.

Peilei Cai, Chengyuan Zhang, and Jian Chai. Forecasting hourly pm2. 5 concentrations based on decomposition-ensemble-reconstruction framework incorporating deep learning algorithms. *Data Science and Management*, 6(1):46–54, 2023.

W Geoffrey Cobourn. An enhanced pm2. 5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. *Atmospheric Environment*, 44(25): 3015–3023, 2010.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019. URL https://arxiv.org/abs/1901.05555.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

ECMWF. *IFS Documentation CY48R1 - Part VIII: Atmospheric Composition*. Number 8. ECMWF, 06/2023 2023. doi: 10.21957/749dc09059.

David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(4):453–476, 1981.

Yufang Hao, Bin Luo, Maimaiti Simayi, Wei Zhang, Yan Jiang, Jiming He, and Shaodong Xie. Spatiotemporal patterns of pm2. 5 elemental composition over china and associated health risks. *Environmental Pollution*, 265:114910, 2020.

Martin Heger, Lukas Vashold, Anabella Palacios, Mala Alahmadi, and Marcelo Acerbi. *Blue skies, blue seas: air pollution, marine plastics, and coastal erosion in the Middle East and North Africa*. World Bank Publications, 2022.

Shenxin Li, Sedra Shafi, Bin Zou, Jing Liu, Ying Xiong, and Bilal Muhammad. Pm2.5 concentration exposure over the belt and road region from 2000 to 2020. *International Journal of Environmental Research and Public Health*, 19(5):2852, March 2022. ISSN 1660-4601. doi: 10.3390/ijerph19052852. URL http://dx.doi.org/10.3390/ijerph19052852.

Xiangdong Li, Ling Jin, and Haidong Kan. Air pollution: a global problem needs local fixes, 2019.

Xiaolan Li, Yanjun Ma, Yangfeng Wang, Ningwei Liu, and Ye Hong. Temporal and spatial analyses of particulate matter (pm 10 and pm 2.5) and its relationship with meteorological parameters over an urban city in northeast china. *Atmospheric Research*, 198:185–193, December 2017. ISSN 0169-8095. doi: 10.1016/j.atmosres.2017.08.023. URL http://dx.doi.org/10.1016/j.atmosres.2017.08.023.

Baolei Lv, W Geoffrey Cobourn, and Yuqi Bai. Development of nonlinear empirical models to forecast daily pm2. 5 and ozone levels in three large chinese cities. *Atmospheric Environment*, 147:209–223, 2016.

Adil Masood, Mohammed Majeed Hameed, Aman Srivastava, Quoc Bao Pham, Kafeel Ahmad, Siti Fatin Mohd Razali, and Souad Ahmad Baowidan. Improving PM2.5 prediction in new delhi using a hybrid extreme learning machine coupled with snake optimization algorithm. *Sci. Rep.*, 13(1): 21057, November 2023.

Muhammad Akhtar Munir, Fahad Shahbaz, and Salman Khan. Efficient localized adaptation of neural weather forecasting: A case study in the mena region. In *NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024. URL https://www.climatechange.ai/papers/neurips2024/42.

Armin Nakhjiri and Ata Abdollahi Kakroodi. Air pollution in industrial clusters: A comprehensive analysis and prediction using multi-source data. *Ecological Informatics*, 80:102504, May 2024. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2024.102504. URL http://dx.doi.org/10.1016/j.ecoinf.2024.102504.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate, 2023a. URL https://arxiv.org/abs/2301.10343.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Sandeep Madireddy, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting, 2023b. URL https://arxiv.org/abs/2312.03876.

Dori Nissenbaum, Ron Sarafian, Yinon Rudich, and Shira Raveh-Rubin. Six types of dust events in eastern mediterranean identified using unsupervised machine-learning classification. *Atmospheric Environment*, 309:119902, September 2023. ISSN 1352-2310. doi: 10.1016/j.atmosenv.2023.119902. URL http://dx.doi.org/10.1016/j.atmosenv.2023.119902.

Narendra Ojha, Amit Sharma, Manish Kumar, Imran Girach, Tabish U Ansari, Som K Sharma, Narendra Singh, Andrea Pozzer, and Sachin S Gunthe. On the widespread enhancement in fine particulate matter across the indo-gangetic plain towards winter. *Scientific reports*, 10(1): 5862, 2020.

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), November 2020. ISSN 1942-2466. doi: 10.1029/2020ms002203. URL http://dx.doi.org/10.1029/2020MS002203.

Ron Sarafian, Dori Nissenbaum, Shira Raveh-Rubin, Vikhyat Agrawal, and Yinon Rudich. Deep multi-task learning for early

warnings of dust events implemented for the middle east. *Npj Clim. Atmos. Sci.*, 6 (1), March 2023.

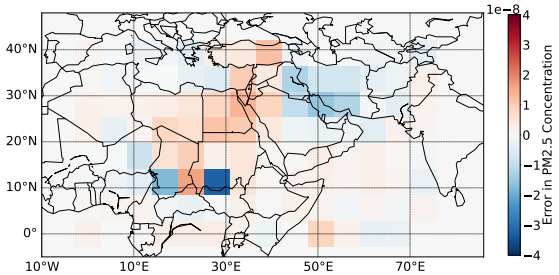
WHO. *Global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, Geneva, 2021. Licence: CC BY-NC-SA 3.0 IGO.

WHO. Ambient (outdoor) air pollution — who.int. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, 2024. [Accessed 20-09-2024].

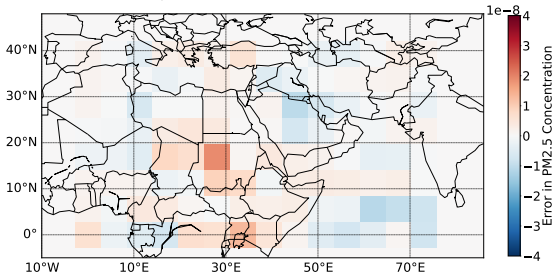
Qianqian Yang, Qiangqiang Yuan, Tongwen Li, Huanfeng Shen, and Liangpei Zhang. The relationships between pm2.5 and meteorological factors in china: Seasonal and regional variations. *International Journal of Environmental Research and Public Health*, 14(12):1510, December 2017. ISSN 1660-4601. doi: 10.3390/ijerph14121510. URL http://dx.doi.org/10.3390/ijerph14121510.

Wenhua Yu, Shanshan Li, Tingting Ye, Rongbin Xu, Jiangning Song, and Yuming Guo. Deep ensemble machine learning framework for the estimation of pm 2.5 concentrations. *Environmental health perspectives*, 130(3): 037004, 2022.

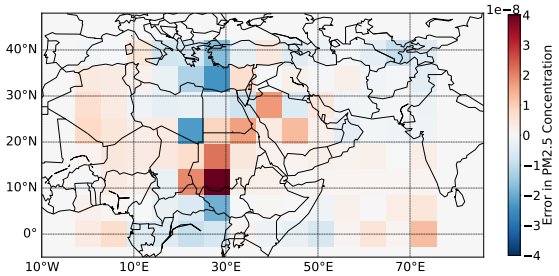
Hongliang Zhang, Jingyi Li, Qi Ying, Jian Zhen Yu, Dui Wu, Yuan Cheng, Kebin He, and Jingkun Jiang. Source apportionment of pm2. 5 nitrate and sulfate in china using a source-oriented chemical transport model. *Atmospheric environment*, 62:228–242, 2012.



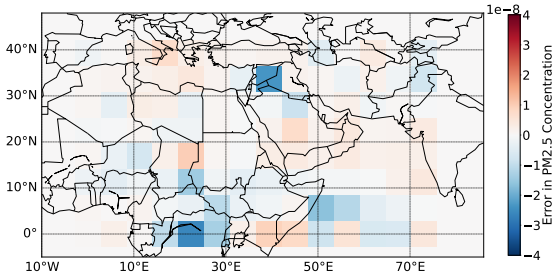
(a) Saudi Arabia – October 29, 2017 (CAMS Forecasts)



(b) Saudi Arabia – October 29, 2017 (Aircast)



(c) Kuwait – October 31, 2017 (CAMS Forecasts)



(d) Kuwait – October 31, 2017 (Aircast)

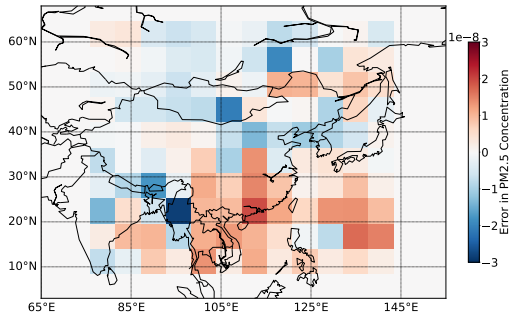
Figure 4: Extreme case visualizations of PM2.5 concentrations (Predictions - Ground Truth) for CAMS global forecasts and Aircast.

Appendix A. Geographic Generalization:

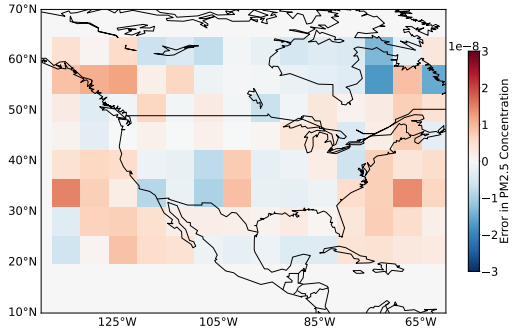
While we only focus the training on the MENA region, we test the geographic generalization ability of our model in East Asia and North America. We find that Aircast slightly over-estimates in some areas (denoted by the red areas in Figure 5).

Table 5: Geographic Generalization: Testing our best model in East Asia and North America with a lead time of 24hrs.

Region	RMSE (μgm^{-3})		
	PM2.5	PM10	PM1
East Asia	10.89	20.61	3.74
North America	9.26	18.22	1.89



(a) East Asia



(b) North America

Figure 5: Extreme case visualizations of PM2.5 concentrations (Predictions - Ground Truth) for Aircast and the CAMS global forecasts.

Appendix B. Varying Seeds

All the ablations in the paper were performed when considering a seed of 42. We additionally test our best setting by varying between 5 different seeds, and report the mean and std. This is a common practice in machine learning research to ensure reproducibility.

Table 6: Varying Seeds for our Best Model. We report the mean, and the standard deviation is reported in the brackets.

Variables	RMSE (μgm^{-3})		
	PM2.5	PM10	PM1
Best Setting	9.00 (0.11)	13.61 (0.20)	6.78 (0.11)

Appendix C. Distribution plots of the PM concentrations

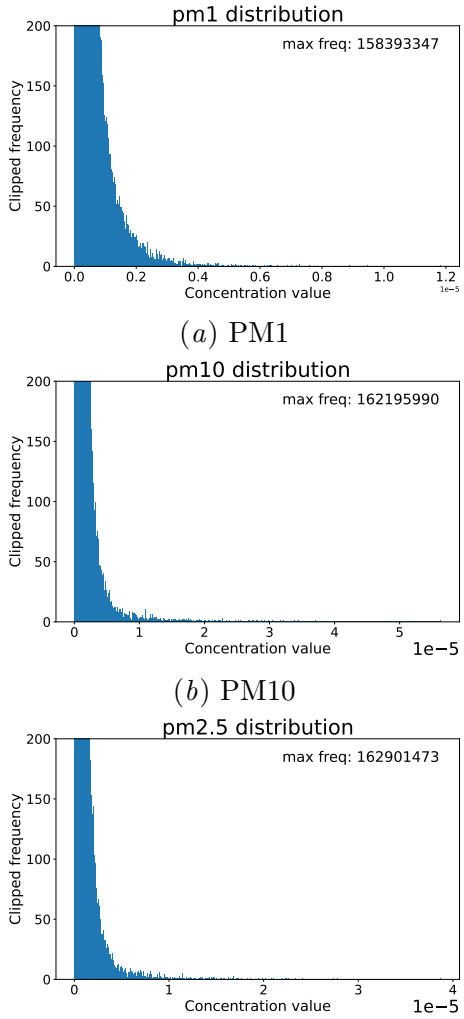
We further show the distribution plots of PM2.5, PM10 and PM1 in Figure 6. Similar to PM2.5, the other 2 concentrations variables PM10 and PM1 also show a long tailed distribution.

Appendix D. Metrics

The **Root Mean Square Error (RMSE)** calculates latitude-weighted ($L(i)$) prediction errors by averaging squared differences between predictions ($\widetilde{X}_{k,i,j}$) and ground truth ($X_{k,i,j}$) across spatial ($H \times W$) and temporal (N) dimensions. The RMSE is calculated as follows,

$$\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} L(i)\left(\widetilde{X}_{k,i,j} - X_{k,i,j}\right)^2\right)} \tag{5}$$

The **Anomaly Correlation Coefficient (ACC)** measures the latitude-weighted ($L(i)$) spatial correlation between predicted ($\widetilde{X}'$)



(a) PM1

(b) PM10

(c) PM2.5

Figure 6: **Skewed distribution of the PM variables.** The x-axis corresponds to the PM variable, and the y-axis corresponds to the frequency clipped at 200 (the maximum frequency is shown in each figure). The clipping is done to visualize the distribution among the low-frequency bins. All the concentration values are in the order of 10^{-5}.

and true (X') anomalies (deviations from climatology C). ACC focuses on pattern similarity rather than absolute values. The ACC is calculated as follows,

$$\frac{\sum_{k,i,j} L(i)\widetilde{X}'_{k,i,j}X'_{k,i,j}}{\sqrt{\left(\sum_{k,i,j} L(i)\widetilde{X}'^2_{k,i,j}\right)\left(\sum_{k,i,j} L(i)X'^2_{k,i,j}\right)}} \tag{6}$$

where anomalies are defined as,

$$\widetilde{X}' = \widetilde{X} - C \quad \text{and} \quad X' = X - C, \tag{7}$$

and the climatology is

$$C = \frac{1}{N}\sum_{k=1}^{N} X_{k,i,j}. \tag{8}$$

Appendix E. Full Results

While the main section of the paper presents metrics only for the three PM variables, we also provide results for all variables using the best-performing configuration in Table 7.

Table 7: Model performance for all variables using the best setting i.e using Surface Weather and Surface Air Quality variables as input to the model. Metrics are RMSE (lower is better) and ACC (higher is better). Lead time is 24hrs.

Variable (unit)	RMSE	ACC
10m U wind (m/s)	2.044	82.7
10m V wind (m/s)	2.248	81.2
2m Temperature (°C)	1.023	92.4
CO (925 hPa) (µg/m^3)	19.4	99.3
Geopotential (925 hPa) (m)	167.121	87.7
O_3 (925 hPa) (µg/m^3)	1.65	97.8
NO_2 (925 hPa) (µg/m^3)	1.75	96.9
NO (925 hPa) (µg/m^3)	1.20	94.9
PM10 (µg/m^3)	13.3	81.4
PM1 (µg/m^3)	6.65	85.3
PM2.5 (µg/m^3)	8.82	82.7
SO_2 (925 hPa) (µg/m^3)	3.41	93.6
Specific humidity (925 hPa) (g/kg)	1.00	79.2
TCNO (mg/m^2)	0.028	100.0
TCCO (mg/m^2)	94.2	71.5
$TCNO_2$ (mg/m^2)	0.384	85.5
$TCSO_2$ (mg/m^2)	0.486	68.0
Temperature (925 hPa) (°C)	1.299	89.5
U wind (925 hPa) (m/s)	2.788	82.4
V wind (925 hPa) (m/s)	2.955	80.3