Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Inverse-Distance Weighting Alignment-DiversityAttention Transformer for Days Ahead Particulate Matter Concentration Forecasting

Yue Yu
Zhejiang University
Pavel Loskot
pavelloskot@intl.zju.edu.cn

ZJU-UIUC Institute
Yu Lin
ZJU-UIUC Institute
Yu Gao
Midea Group (China)

Article

Keywords:

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10013387/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 5th, 2026.
See the published version at https://doi.org/10.1038/s41598-026-71472-8.

An Inverse-Distance Weighting Alignment-Diversity Attention Transformer for Days Ahead Particulate Matter Concentration Forecasting

Yue Yu[1], Pavel Loskot[2,*], Yu Lin[2], and Yu Gao[3]

[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310058, China
[2]Zhejiang University - University of Illinois Urbana-Champaign Institute, Haining 314400, China
[3]AI Research Center, Midea Group, Shanghai 201103, China
[*]pavelloskot@intl.zju.edu.cn

ABSTRACT

An accurate forecasting of the particulate matter (PM) concentrations including $PM_{2.5}$ and PM_{10} concentrations is vital for the public health and the environment management. Forecasting highly non-stationary PM time series is challenging due to their complex spatial and temporal dependencies and irregular local correlations. In this paper, an adaptive Inverse-Distance Weighting (IDW) iTransformer-based framework referred to as IDW-ADAFormer is devised for the medium-term joint forecasting of spatially distributed multi-sensor PM data. In particular, the adaptive IDW decomposition first separates the spatial trends from the local temporal variations. The components are aligned by a newly constructed kernel-based alignment diversity attention (ADA) module. The aligned components and the raw input data are then adaptively fused using learnable weights. The model training minimizes a newly constructed multi-objective AQI-aware loss function, which accounts for the prediction accuracy, the trend preservation, and the alignment with the Air-Quality Indicator (AQI) levels. The numerical experiments were performed for a public dataset of hourly $PM_{2.5}$ and PM_{10} data. Assuming one to three days ahead predictions, the proposed IDW-ADAFormer was found to consistently outperform other existing methods, while maintaining the interpretability of an iTransformer.

Introduction

The environmental control often dictates the mass deployments of various types of sensors[1]. For example, the $PM_{2.5}$ and PM_{10} data must be monitored to limit unsafe concentrations of pollutants, which were shown in many studies to be dangerous to human health[2–4]. Epidemiological studies link even short-term exposures to these pollutants with a higher risk of respiratory and cardiovascular diseases[5–7]. The public health advisories require predicting the future values of pollutants in order to allow taking precautionary steps in time. The accurate pollutant forecasting is therefore essential for public health and environmental management, and for implementing effective mitigation strategies.

The $PM_{2.5}$ and PM_{10} data are typically collected from a network of spatially distributed sensors. These data experience intricate spatial and temporal dependencies. The temporal variations can be described by extracting the trend and the seasonality components, although the underlying time series are generally non-stationary. The common decomposition methods include Empirical Mode Decomposition (EMD), Variational Mode Decomposition (VMD), the Seasonal-Trend Decomposition (STD), and a Wavelet Transform (WT). The decompositions can stabilize the model learning without sacrificing the model interpretability. The single-sensor AQI data are forecasted using the EMD followed by a bi-directional long-short term memory (BiLSTM) model in[8]. The VMD is combined with deep learning to obtain a stable and noise-resistant forecasting of AQI data in[9]. The $PM_{2.5}$ values are predicted using the STD with a gated recurrent unit (GRU) in[10], and using the discrete WT followed by a transformer in[11], respectively. However, these forecasting architectures have several limitations. First, they typically consider only temporal variations, and neglect the sample dependencies across multiple sensors. Second, adaptive decomposition methods such as EMD are prone to information leakage and semantic inconsistencies[9,10]. Such inconsistencies arise when the decompositions are assumed independently over segments. For example, the signal envelopes in EMD obtained for different segments are not semantically aligned. Third, the decompositions increase the overall computational complexity, and the model training costs.

The complex space-time dependencies in multi-sensor data can be modeled using the attention in transformers[12–14]. For example, the iTransformer treats each time series as a variate token, and its attention evaluates the inter-time series correlations[15]. The CrossFormer improves the modeling effectiveness by employing a two-stage attention in order to separate the modeling of spatial and temporal correlations[16]. The ETSformer captures the temporal lags and periodicity of time series using a frequency

domain self-attention with exponential smoothing[17]. The recent innovations include a two-stage attention performed across temporal segments and hidden layers[18]. A dual attention with multi-scale adaptive fusion over input segments was investigated in[19]. A multi-resolution transformer was devised in[20] to capture the component dependencies at multiple frequencies. However, the standard attention mechanism employed in these methods biases the modeling towards a few dominant long-term patterns, while more complex local patterns are averaged out.

The majority of existing models for air pollutant forecasting rely on a single numerical objective such as the Mean-Square Error (MSE)[21,22], and the Root MSE (RMSE)[23]. The objective function can be modified by considering weighted prediction errors as demonstrated in[24]. The Chi-square divergence regularization is adopted in[25], whereas the L_2 regularization is assumed in[26] to reduce the model overfitting. However, these loss functions do not explicitly consider alignment with the defined AQI levels, which limits interpretability and utility of the predictions.

This paper addresses some of these challenges by introducing an iTransformer-based forecasting framework for multi-sensor PM data. The proposed architecture is referred to as IDW-ADAFormer. In particular, the input time series are first segmented, and decomposed before the components are embedded by a modified attention module having two kernels. The resulting spatial-temporal embeddings are optimally combined using learnable weights. Finally, the subsequent iTransformer generates the predicted samples. The model is trained assuming a multi-objective loss function. The specific contributions reported in this paper can be summarized as follows.

1. An adaptive k-nearest IDW decomposition is introduced to separate the spatial trends from the local temporal variations. Such a decomposition does not cause information leakage, since it can be performed independently at each time step. In addition and importantly, this decomposition does not contain any trainable parameters, so it does not affect the model training.

2. An alignment-diversity attention (ADA) module is introduced to combine the raw input samples with the embedded features. It allows controlling the degree of feature embedding before the subsequent processing. The ADA module exploits two kernels. The Cross-sensor Alignment Kernel (CAK) captures the spatial patterns across sensors, whereas the Diversity-Correlation Kernel (DCK) models complex temporal dependencies. Such a design enables more precise local modeling of complex temporal dependencies and variations.

3. A multi-objective AQI-aware loss function is designed to jointly account for the prediction accuracy, trend preservation, and consistency of the AQI levels. It boosts the prediction performance as well as enhances the model interpretability.

4. The numerical experiments show that the proposed architecture outperforms other existing methods for one to three days ahead forecasting of the PM data, while offering the comparable robustness and efficiency as the existing methods.

The remainder of this paper is organized as follows. Section provides a detailed description of the proposed IDW-ADAFormer architecture. Section presents numerical results to evaluate the accuracy and efficiency of the proposed architecture. The proposed architecture and the results are discussed in Section . Section concludes the paper, and outlines suggestions for future work.

Methodology

The proposed IDW-ADAFormer is a multi-input multi-output (MIMO) architecture. It is designed to allow predicting the future values of multivariate time series generated by N spatially separated sensors. The past L_p observations representing a lookback window are arranged in a matrix, $\mathbf{X}_{1:L_p} \in \mathbb{R}^{N\times L_p}$. The objective is to predict the next L_h samples within the prediction horizon, i.e.,

$$\hat{\mathbf{Y}}_{L_p+1:L_p+L_h} = f_\Theta\left(\mathbf{X}_{1:L_p}\right) \in \mathbb{R}^{N\times L_h}, \tag{1}$$

where θ are learnable parameters of the predictor, f. The parameters, θ, are determined in order to minimize the loss function, $\mathcal{L}$, i.e.,

$$\Theta^* = \arg\min_\Theta \mathcal{L}\left(\hat{\mathbf{Y}}_{L_p+1:L_p+L_h}, \mathbf{X}_{L_p+1:L_p+L_h}\right). \tag{2}$$

The overall architecture of the proposed IDW-ADAFormer is shown in Figure 1. It focuses on preprocessing the input samples to allow more effectively exploiting the spatial sample dependencies, the temporal dynamics, and the AQI-levels semantic by the subsequent iTransformer. In particular, the input multivariate time series are first segmented, and normalized using a Reversible Instance Normalization (RevIN). The adaptive k-nearest IDW decomposition module extracts a spatially-interpolated

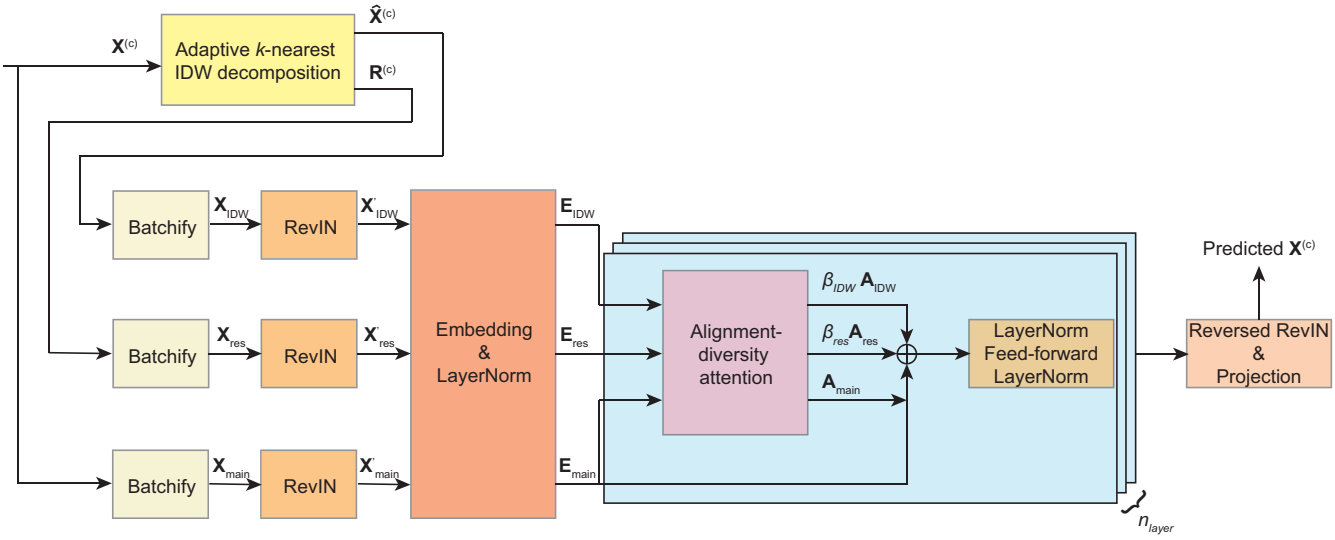


Figure 1. A block diagram of the proposed IDW-ADAFormer consists of the adaptive k-nearest IDW decomposition module and alignment-diversity attention module followed by iTransformer to generate predictions.

component capturing the spatial trends and a residual component, which preserves the local temporal variations. These components together with the original input samples are jointly embedded into a shared vector space, and then fed into the proposed ADA module. The ADA module employs two kernels. The CAK kernel aligns the spatial patterns among time series from different sensors. The DCK kernel models the heterogeneous temporal dependencies. Both kernel outputs are fused using learnable weights, which enables optimally balancing the output contributions to the final predictions. The combined outputs are processed by another feedforward layer, and the reverse RevIN is used to restore the original range of values. The final predictions are generated using a projection head within iTransformer. The IDW-ADAFormer training is guided by a multi-objective AQI-aware loss function. It minimizes the prediction errors, preserves the temporal trends, and also enforces the consistency of predicted samples with defined AQI-levels.

Adaptive k-nearest IDW decomposition

The adaptive k-nearest IDW decomposition leverages the spatial locations of sensors by decomposing the input samples into a IDW-estimated component and a residual component, respectively. Unlike other decomposition methods that must be performed over whole segments of samples, the proposed method enables a sample-wise decomposition. It preserves the signal resolution, and also prevents information leakage. Recall that the last L_p measurements from N sensors are denoted as $\mathbf{X}_{1:L_p}$. The pairwise spatial distances, d_{ij}, between sensors are computed using the Haversine formula[27]. The distances are used to partition the sensors into overlapping subsets, $\mathcal{N}_{s,t}^{(k)}$, each containing the k nearest neighbors. Furthermore, assuming the IDW spatial interpolation, the pollutant concentrations, $x_{s,t} \in \mathbf{X}_{1:L_p}$, corresponding to sensor s at time t can be estimated as,

$$\hat{x}_{s,t} = \frac{\sum_{j \in \mathcal{N}_{s,t}^{(k)}} d_{s,j}^{-2} x_{j,t}}{\sum_{j \in \mathcal{N}_{s,t}^{k}} d_{s,j}^{-2}}. \tag{3}$$

The IDW estimate (3) captures the spatially shared component of the pollutant concentration, and the residual component represents the local variations that cannot be explained by the data from neighboring sensors, i.e.,

$$x_{s,t} = \hat{x}_{s,t} + x_{s,t}^{\text{res}}, \quad 1 \le s \le N,\ 1 \le t \le L_p. \tag{4}$$

In matrix form, the IDW decomposition is written as,

$$\mathbf{X}_{\text{main}} = \hat{\mathbf{X}}_{\text{IDW}} + \mathbf{X}_{\text{res}}. \tag{5}$$

Both components in (5) are used as the features in subsequent processing. Furthermore, in order to account for the dynamically varying spatial correlations, the neighborhood size k is adaptively selected for each sensor and time step to minimize the reconstruction error of the IDW estimate, i.e.,

$$k_{s,t}^{*} = \arg\min_{k} \left| x_{s,t} - \hat{x}_{s,t}^{(k)} \right|^2. \tag{6}$$

IDW-ADAFormer

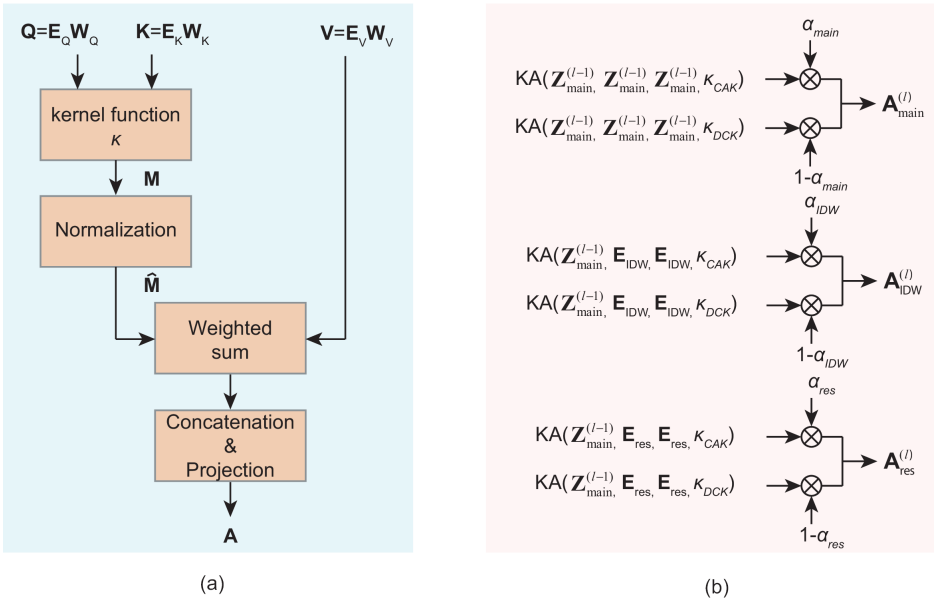
The ADA module fuses the original measurements, $\mathbf{X}_{\text{main}}$, the IDW estimates, $\mathbf{X}_{\text{IDW}}$, and the residuals, $\mathbf{X}_{\text{res}}$, before generating the predicted samples, $\hat{\mathbf{Y}}$. The data processing is effectively performed in batches of B matrices. In particular, the data sample from sensor s at time t and batch b is first normalized using the reversible RevIN as,

$$x'_{b,s,t} = \gamma_s \cdot \frac{x_{b,s,t} - \mu_{b,s}}{\sqrt{\sigma^2_{b,s} + \epsilon}} + \beta_s, \tag{7}$$

where $\mu_{b,s}$ is the sample mean, and $\sigma_{b,s}$ is the sample variance, respectively, i.e.,

$$\mu_{b,s} = \frac{1}{L_p}\sum_{t=1}^{L_p} x_{b,s,t}, \text{ and, } \sigma_{b,s} = \sqrt{\frac{1}{L_p}\sum_{t=1}^{L_p}(x_{b,s,t} - \mu_{b,s})^2}. \tag{8}$$

The normalized sequences of samples are linearly projected into the embeddings, $\mathbf{E}_{\text{main}}$, $\mathbf{E}_{\text{IDW}}$, and $\mathbf{E}_{\text{res}}$, using the same linear transformation and the bias. The embeddings are further normalized using the standard LayerNorm to stabilize the training. The normalized embeddings are processed by n_{layer} layers of the ADA module and the output feed-forward network. Both CAK and DCK kernels are implemented as a generic attention mechanism involving three defined linear transformations, $\text{KA}(\mathbf{E}_Q, \mathbf{E}_K, \mathbf{E}_V, \kappa)$, and a non-linearity, κ. A general attention mechanism is depicted in Figure 2(a), whereas the proposed modified ADA mechanism is shown in Figure 2(b).



(a) (b)

Figure 2. Block diagrams of (a) the canonical kernel attention, and (b) the proposed alignment-diversity attention.

Denote as $\mathbf{Z}^{(l)}_{\text{main}}$ the output of the l-th layer of the ADA module, and let, $\mathbf{Z}^{(0)}_{\text{main}} = \mathbf{E}_{\text{main}}$. The attention outputs are written as,

$$\begin{aligned}
\mathbf{A}^{(l)}_{\text{main}} &= \alpha_{\text{main}}\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{Z}^{(l-1)}_{\text{main}}, \kappa_{\text{CAK}}) + (1-\alpha_{\text{main}})\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{Z}^{(l-1)}_{\text{main}}, \kappa_{\text{DCK}}),\\
\mathbf{A}^{(l)}_{\text{IDW}} &= \alpha_{\text{IDW}}\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{E}_{\text{IDW}}, \mathbf{E}_{\text{IDW}}, \kappa_{\text{CAK}}) + (1-\alpha_{\text{IDW}})\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{E}_{\text{IDW}}, \mathbf{E}_{\text{IDW}}, \kappa_{\text{DCK}}),\\
\mathbf{A}^{(l)}_{\text{res}} &= \alpha_{\text{res}}\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{E}_{\text{res}}, \mathbf{E}_{\text{res}}, \kappa_{\text{CAK}}) + (1-\alpha_{\text{res}})\,\text{KA}(\mathbf{Z}^{(l-1)}_{\text{main}}, \mathbf{E}_{\text{res}}, \mathbf{E}_{\text{res}}, \kappa_{\text{DCK}}).
\end{aligned} \tag{9}$$

The three components (9) are then split for processing by n_{head} attention heads of dimension, $d_{\text{head}} = D/n_{\text{head}}$, each. The corresponding linear projections form the query, key and value matrices, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times n_{\text{head}}\times N\times d_{\text{head}}}$, respectively. The attention heads compute similarity matrices, $\mathbf{M}_{b,h,i,j} = \kappa(\mathbf{q}_{b,h,i}, \mathbf{k}_{b,h,j})$, where $\mathbf{q}_{b,h,i}, \mathbf{k}_{b,h,j} \in \mathbb{R}^{d_{\text{head}}}$, which are then normalized over the keys, i.e.,

$$\hat{\mathbf{M}}_{b,h,i,j} = \frac{\mathbf{M}_{b,h,i,j}}{\sum_j \mathbf{M}_{b,h,i,j} + \epsilon}. \tag{10}$$

The attention outputs are linearly weighted values, $\mathbf{V}$, such that,

$$\mathbf{A}_{b,h,i} = \sum_j \hat{\mathbf{M}}_{b,h,i,j} \mathbf{V}_{b,h,j}. \tag{11}$$

The proposed ADA module shown in Figure 2(b) utilizes two non-linear kernel functions. In particular, the CAK function, κ_{CAK}, generates an inductive bias encouraging the attention alignment across sensors. It is defined as,

$$\kappa_{\text{CAK}}(\mathbf{q}, \mathbf{k}) = \exp\left(- \frac{2\sin^2\left(\pi(n_q - n_k)/p\right)}{\ell^2} \right), \tag{12}$$

where n_q and n_k are the variate indices associated with the query and the key, respectively. The parameter, p, is a learnable period controlling the alignment frequency. The parameter, ℓ, is a learnable length-scale controlling the kernel smoothness. In addition, the DCK function linearly combines the Gaussian, Laplace, and Cauchy components to allow modeling diverse cross-variate dependencies, i.e.,

$$\kappa_{\text{DCK}}(\mathbf{q}, \mathbf{k}) = \alpha_G \exp\left(- \frac{\|\mathbf{q} - \mathbf{k}\|^2}{2\ell_G^2} \right) + \alpha_L \exp\left(- \frac{\|\mathbf{q} - \mathbf{k}\|}{\gamma_L} \right) + \alpha_C \frac{1}{1 + (\|\mathbf{q} - \mathbf{k}\| / b_C)^2}, \tag{13}$$

where $\|\cdot\|$ denotes the Euclidean norm. The parameters, α_G, α_L, and, $\alpha_C \geq 0$, are the learnable weights. The parameters, ℓ_G, γ_L, and, b_C, are the learnable scaling factors. The kernel definition (13) can adaptively model the heterogeneous dependencies among different types of random variables. Supplementary Figures S1 and S2 illustrate the shapes of the kernel functions (12) and (13), respectively. More importantly, note that the non-linear kernels (12) and (13) replace the dot-product distances in the original transformer attention. It enables the ADA module to perform the cross-sensor alignment across diverse correlation patterns. Supplementary Table S1 indicates that the computational complexity of the ADA module is, $O(BN^2D)$ (Big-O notation), which is comparable to the complexity of the standard attention.

The embeddings from the attention layers are fused together using learnable weights, normalized, and then processed by another feed-forward network having the Gaussian Error Linear Unit (GELU) activation and a dropout. Specifically, the output of the l-th layer, $1 \leq l \leq n_{\text{layer}}$, is written as,

$$\mathbf{Z}_{\text{main}}^{(l)} = \text{LN}\left(\text{FF}\left(\text{LN}\left(\mathbf{Z}_{\text{main}}^{(l-1)} + \mathbf{A}_{\text{main}}^{(l)} + \beta_{\text{IDW}} \cdot \mathbf{A}_{\text{IDW}}^{(l)} + \beta_{\text{res}} \cdot \mathbf{A}_{\text{res}}^{(l)} \right) \right) \right). \tag{14}$$

At the output of the n_{layer}-th layer, the reversed RevIN restores the original range of values, i.e.,

$$\hat{x}_{b,s,t} = \frac{\hat{x}'_{b,s,t} - \beta_s}{\gamma_s} \sqrt{\sigma_{b,s}^2 + \epsilon} + \mu_{b,s}. \tag{15}$$

Finally, the forecasted values are obtained using another linear projection.

Multi-objective AQI-aware loss

As explained above, it is desirable to guide the training of the IDW-ADAFormer by adopting a multi-objective loss function. The proposed loss function combines three components, and it is inspired by uncertainty-weighted multiple objectives introduced in[28].

The MSE loss quantifies the standard prediction error as,

$$\mathcal{L}_1 = \frac{1}{BL_hN} \sum_{b,t,s} (\hat{y}_{b,t,s} - y_{b,t,s})^2, \tag{16}$$

where $\hat{y}_{b,t,s}$ and $y_{b,t,s}$ are the predicted and the true pollutant concentrations, respectively.

The shape loss is introduced to encourage the temporal consistency of predicted samples. It is defined as,

$$\mathcal{L}_2 = \frac{w_{\text{shape}}}{B(L_h - 1)N} \sum_{b,t,s} |(\hat{y}_{b,t+1,s} - \hat{y}_{b,t,s}) - (y_{b,t+1,s} - y_{b,t,s})|, \tag{17}$$

where w_{shape} denotes a learnable weight, which controls the shape consistency.

The AQI-level loss enforces the agreed AQI categories as specified by the U.S. EPA standard[29]. The thresholds defining these categories of the PM concentrations averaged over 24 hours are summarized in Supplementary Table S2. Thus, the true

and the predicted PM concentrations, $\bar{\hat{y}}$, and, $\bar{y}$, in batch, b, as measured by sensor, s, at time, t, are averaged over 24-hour windows. The averaged values are assigned the probabilities that they belong to one of the $k = 1, 2, \ldots, 6$, defined AQI-levels as,

$$p^{\text{pred}}_{b,t,s,k} = \frac{\exp(-|\bar{\hat{y}}_{b,t,s} - \ell_k|)}{\sum_{k'} \exp(-|\bar{\hat{y}}_{b,t,s} - \ell_{k'}|)}, \quad p^{\text{true}}_{b,t,s,k} = \frac{\exp(-|\bar{y}_{b,t,s} - \ell_k|)}{\sum_{k'} \exp(-|\bar{y}_{b,t,s} - \ell_{k'}|)}, \tag{18}$$

where ℓ_k denotes the k-th AQI lower-bound threshold as defined in Supplementary Table S2. The calculated probabilities (18) act as the soft assignment to the defined AQI levels. These probabilities can be further optimized using stable gradients across the AQI level boundaries. The resulting total AQI-level loss is computed as,

$$\mathcal{L}_3 = \frac{w_{\text{AQI}}}{BL'_hN} \sum_{b,t,s,k} |p^{\text{true}}_{b,t,s,k} - p^{\text{pred}}_{b,t,s,k}|, \tag{19}$$

where w_{AQI} is a learnable weight, and $L'_h = (L_h - w + 1)$ is adjusted length of the predicted horizon.

Finally, the three losses are linearly combined using adjustable weights, $w_i^{(\mathcal{L})} = (2\exp(\log \sigma_i^2))^{-1}$, as,

$$\mathcal{L} = w_1^{(\mathcal{L})}\mathcal{L}_1 + w_2^{(\mathcal{L})}\mathcal{L}_2 + w_3^{(\mathcal{L})}\mathcal{L}_3. \tag{20}$$

Results

Numerical results assume the time series data of $PM_{2.5}$ and PM_{10} concentrations. The forecasting of the PM concentrations for one to three days ahead are performed by the proposed IDW-ADAFormer and several other existing frameworks that were proposed in the literature. The forecasting accuracy is evaluated using several performance metrics. In addition, the frameworks are also compared in terms of their computational complexity.

Dataset description

The PurpleAirSF dataset[30] contains the air quality measurements collected from a network of low-cost spatially distributed sensors in San Francisco bay area in the USA. This dataset contains the measurements for several air pollutants; however, only the $PM_{2.5}$ and PM_{10} concentrations are considered in our forecasting experiments, since these pollutants are known to have the most significant effect on the human health. The measurements from $N = 14$ sensors at one hour resolution over 568 consecutive days between October 25, 2021 and May 15, 2023 were selected for our study. The statistical properties of these time series data are summarized in supplementary Section S3.

Evaluation metrics

Recall that $\mathbf{Y}_{L_p+1:L_p+L_h} \in \mathbb{R}^{N\times L_h}$, and $\hat{\mathbf{Y}}_{L_p+1:L_p+L_h} \in \mathbb{R}^{N\times L_h}$ denote the true and the predicted pollutant concentrations over the prediction horizon of L_h samples, respectively. The following four metrics are used to evaluate the performance.

The Mean Absolute Error (MAE) is defined as,

$$\text{MAE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{NL_h} \sum_{s=1}^{N} \sum_{t=1}^{L_h} \left|Y_{s,t} - \hat{Y}_{s,t}\right|. \tag{21}$$

The Root Mean Square Error (RMSE) is defined as,

$$\text{RMSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \sqrt{\frac{1}{NL_h} \sum_{s=1}^{N} \sum_{t=1}^{L_h} \left(Y_{s,t} - \hat{Y}_{s,t}\right)^2}. \tag{22}$$

The symmetric Mean Absolute Percentage Error (sMAPE) is defined as,

$$\text{sMAPE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{NL_h} \sum_{s=1}^{N} \sum_{t=1}^{L_h} \frac{\left|Y_{s,t} - \hat{Y}_{s,t}\right|}{(|Y_{s,t}| + |\hat{Y}_{s,t}|)/2} \times 100\% \,. \tag{23}$$

In addition, the AQI-level accuracy evaluates whether the predicted pollutant concentrations fall within the correct AQI categories as specified by the U.S. EPA standard[29]. Specifically, a 24-hour sliding window average of the pollutant concentrations is computed as,

$$\bar{Y}_{s,t} = \frac{1}{24} \sum_{i=1}^{24} Y_{s,t+i}, \quad \bar{\hat{Y}}_{s,t} = \frac{1}{24} \sum_{i=1}^{24} \hat{Y}_{s,t+i}. \tag{24}$$

The average concentration values are then mapped to the defined AQI categories. The corresponding AQI-level accuracy is defined as,

$$\text{AQI-Acc} = \frac{1}{NT'} \sum_{s=1}^{N} \sum_{t=1}^{T'} I\Big(\Phi(\hat{\bar{Y}}_{s,t}), \Phi(\bar{Y}_{s,t})\Big), \tag{25}$$

where $\Phi(\cdot)$ is the pollutant concentration quantized to the AQI levels as defined by the thresholds in supplementary Table S2, and, $T' = L_h - 24$, is the effective horizon length after applying the 24-hour sliding window averaging. The indicator function, $I(x, y) = 1$, if $x = y$, and $I(x, y) = 0$, if $x \neq y$.

Experimental setting

The following state-of-the art (SOTA) forecasting architectures are considered for comparison: iTransformer[15], CrossFormer[16], DLinear[31], FreTS[32], TimeMixer[33], ETSformer[17], BigST[34], and PatchSTG[35]. These models represent a wide range of diverse forecasting approaches and architectures for time series data, and thus, comparing their performances has a value in itself. These architectures are summarized in supplementary Section S3.

All experiments were carried out on an NVIDIA GeForce RTX 4060 GPU. The batch size was set to 128, and 150 training epochs were used. The initial learning rate of the Adam optimizer was set to 10^{-3}. The L_p historical hourly values are used to forecast the next L_h hourly values of the pollutant concentrations. The three pairs of lookback and forecasting horizon lengths, (L_p, L_h), were evaluated, i.e., $(72, 24)$, $(96, 48)$, and $(120, 72)$, respectively. The time series data were split chronologically into the training, validation, and testing samples at a ratio of $8 : 1 : 1$. The validation set is used for the model selection, and hyperparameter tuning. The unseen samples in test set are used to evaluate the model performance.

The optimal embedding and hidden dimensions are chosen among the values, $\{16, 32, 64, 128\}$. The optimal number of attention heads is chosen among the values, $\{2, 4, 8\}$. The optimal number of layers is chosen among the values, $\{1, 2, 3\}$. The hyperparameter optimization is performed using the automated framework Optuna[36]. The models are trained assuming two loss functions, i.e., a simple MSE-only loss, and the multi-objective loss defined earlier in (20). The model training adopts an early stopping with the patience value of 5, so that the best-performing model configuration on the validation set can be selected.

Comparison results

Tables 1–4 compare the performances of different prediction architectures for the $PM_{2.5}$ and PM_{10} datasets assuming the two loss functions. These results suggest that the proposed IDW-ADAFormer consistently outperforms or is among the best performing architectures in all scenarios considered. Due to a space limit, the results of ablation experiments are presented in supplementary Section S3. The ablation results show that all three components in the proposed attention mechanism are important for the balanced and robust performance. The results shown in supplementary Section S3 include the 95% confidence interval half-widths for different baseline models, the row-normalized confusion matrix of misclassified samples of the proposed model, the optimal k values, and the computational complexity of all models considered in terms of the required floating-point operations (FLOPs) and the number of model parameters.

Table 1 reports the performances for predicting the $PM_{2.5}$ values under the MSE loss. The proposed IDW-ADAFormer achieves the lowest MAE, RMSE and sMAPE values for almost all forecasting settings. In some cases, the iTransformer achieves marginally better AQI-Acc than the IDW-ADAFormer. Other models such as DLinear and FreTS have competitive AQI-Acc performance, but their performance is inferior in other performance metrics. The CrossFormer, TimeMixer, and ETSformer exhibit larger prediction errors as well as lower AQI-level classification accuracy, so they are less suited for the $PM_{2.5}$ predictions under the MSE loss. Furthermore, we compared the performance of BigST and PatchSTG, which are the models that explicitly incorporate the spatial dependencies among sensor data. It was observed that BigST achieves a moderate forecasting performance, and PatchSTG performs less competitively on $PM_{2.5}$ forecasting. The latter may be attributed to the patch-based spatial representations that are effective for periodic and structured patterns, but not irregular patterns found in the PM data.

Table 2 reports the same results as Table 1, but assuming the multi-objective AQI-aware loss instead. It can be observed that most performance metrics benefit from modifying the target loss function and incorporating AQI-level information. Overall, the improvement is most pronounced in MAE and sMAPE. Under the three settings $(L_p, L_h) = (72, 24)$, $(96, 48)$, and $(120, 72)$, MAE is reduced by 2.4315%, 1.3712%, and 2.8533%, respectively, on average for all models compared with the MSE loss. The corresponding sMAPE is reduced by 2.1347%, 0.6159%, and 2.0311%, respectively. The limited improvement in RMSE may be attributed to the fact that the training objective already optimizes the squared error, which strongly penalizes large deviations, which leaves less room for any additional gains from other objectives. The AQI-level accuracy shows a marginal improvement. This may be due to coarse and discontinuous nature of the AQI thresholds, where the improved point-wise accuracy does not necessarily lead to boundary-level changes in the discrete AQI categories. More importantly, the IDW-ADAFormer again achieves the best values in MAE, RMSE, and sMAPE for all prediction horizons considered, so it is robust even when more

Table 1. The forecasting performance of different models for $PM_{2.5}$ time series under the MSE loss.

Method	$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
IDW-ADAFormer	**1.8741**	**5.9057**	0.9735	**78.4208**	**1.9871**	**6.0093**	**0.9751**	**81.1905**	**2.0440**	**6.0631**	**0.9762**	**83.0603**
iTransformer	1.8956	5.9371	**0.9737**	78.8347	2.0281	6.0288	0.9750	82.4056	2.0462	6.0743	0.9752	83.5523
CrossFormer	3.4883	7.5822	0.8343	109.0922	3.8744	7.8662	0.8511	111.2984	4.0021	7.9279	0.8035	115.6712
DLinear	2.0036	6.0468	0.9708	81.4457	2.1161	6.0840	0.9736	85.1867	2.1635	6.1163	0.9758	86.8433
FreTS	1.9870	5.9459	0.9724	83.3550	2.2115	6.0290	0.9747	89.2772	2.3608	6.1135	0.9753	92.7535
TimeMixer	2.4891	6.2243	0.9408	94.7763	2.7825	6.4943	0.9642	98.0133	2.8406	6.5540	0.9645	97.2928
ETSformer	2.1144	6.0047	0.8779	95.3897	2.2432	6.1242	0.8633	99.6001	2.3194	6.1985	0.8620	101.7704
BigST	2.1616	5.9289	0.9733	86.3820	2.5198	6.0812	0.9706	95.4268	2.7336	6.1820	0.9712	99.9787
PatchSTG	2.7331	6.6521	0.9452	99.2280	3.4036	7.0767	0.9282	107.5759	3.9122	7.5073	0.8926	117.9742

complex optimization objectives are adopted. The results in Table 2 also indicate that the IDW-ADAFormer performs well in terms of both the absolute and relative prediction errors. The iTransformer remains the closest competitor, and it has a slightly larger AQI-Acc when shorter horizon windows are assumed. Both DLinear and FreTS achieve a good performance, which is, however, not as good as the iTransformer-based methods. The CrossFormer exhibits large sMAPE, and much smaller AQI-Acc. The TimeMixer and ETSformer underperform in terms of prediction accuracy and the AQI-level classification.

Table 2. The forecasting performance of different models for $PM_{2.5}$ time series under the AQI-aware loss.

Method	$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
IDW-ADAFormer	**1.8610**	**5.8997**	0.9737	**77.8353**	**1.9863**	**6.0054**	**0.9751**	**80.9647**	**2.0333**	**6.0554**	**0.9763**	**82.7519**
iTransformer	1.8677	5.9157	**0.9742**	77.8963	2.0237	6.0251	**0.9751**	82.3214	2.0508	6.0668	0.9757	83.2042
CrossFormer	3.2976	6.9717	0.9324	101.6701	3.7274	7.8210	0.8561	111.3024	3.9189	7.7809	0.8011	114.5989
DLinear	1.9926	6.0511	0.9708	81.0526	2.1095	6.0794	0.9736	85.0564	2.1423	6.1040	0.9758	86.3236
FreTS	1.9468	5.9459	0.9724	82.2756	2.1076	6.0272	0.9742	86.5635	2.1777	6.1174	0.9759	88.1106
TimeMixer	2.2932	6.1897	0.9410	90.1459	2.7820	6.4939	0.9642	98.0009	2.7118	6.5168	0.9645	94.4923
ETSformer	2.0853	5.9867	0.8846	93.7519	2.2282	6.1103	0.8683	98.8328	2.2998	6.1907	0.8630	101.0083
BigST	2.1129	5.9182	0.9733	84.6746	2.4936	6.0737	0.9711	94.9066	2.6331	6.1543	0.9723	98.1341
PatchSTG	2.7299	6.6520	0.9454	99.1531	3.3512	7.0463	0.9298	106.8891	3.6966	7.2989	0.9133	111.4988

Table 3 reports the same results as Table 1, but assuming the PM_{10} data. The proposed IDW-ADAFormer consistently achieves the best MAE, RMSE, and sMAPE values for all three prediction horizons considered. The AQI-Acc values are very close to an ideal 100% for the IDW-ADAFormer, DLinear, and FreTS. It implies that the classification of AQI levels is easier for the PM_{10} data than for the $PM_{2.5}$ data. Even though the FreTS have a slightly better AQI-Acc for some prediction horizons, it suffers from noticeably larger values of MAE and sMAPE. Overall, a relative ranking of different methods is similar to the one observed for the $PM_{2.5}$ data.

Table 3. The forecasting performance of different models for PM_{10} time series under the MSE loss.

Method	$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
IDW-ADAFormer	**2.2072**	**7.0174**	**0.9997**	**70.2151**	**2.3354**	**7.1237**	0.9998	**73.1807**	**2.3740**	**7.1916**	0.9994	**74.7855**
iTransformer	2.2509	7.0523	0.9994	70.9301	2.3392	7.1265	0.9999	73.5884	2.3903	7.1959	0.9999	74.7960
CrossFormer	3.9941	8.8579	0.9632	102.3338	5.2701	9.3915	0.9956	111.9644	5.1546	9.5849	0.9967	110.0761
DLinear	2.3389	7.1466	0.9994	73.5959	2.4515	7.1935	0.9998	77.2156	2.4914	7.2246	0.9999	78.5853
FreTS	2.4297	7.0456	0.9996	79.2319	2.5438	7.1304	**1.0000**	81.4278	2.6658	7.2090	0.9999	84.4484
TimeMixer	2.8316	7.3371	0.9849	86.3149	3.4955	7.7105	0.9930	96.2188	3.2669	7.7048	**1.0000**	89.6509
ETSformer	2.4743	7.0989	0.9235	86.6670	2.6054	7.2123	0.9038	91.8488	2.7122	7.3290	0.9088	93.9044
BigST	2.5169	7.0260	0.9993	77.9804	3.0119	7.1880	0.9999	90.2208	3.3576	7.3493	0.9999	95.8057
PatchSTG	3.4747	8.3018	0.9894	94.3313	3.9717	8.9714	0.9752	101.9971	4.1933	9.0378	0.9641	105.3483

Table 4 reports the same results as Table 1, but assuming the PM_{10} data, and assuming the multi-objective AQI-aware loss. It can be observed that most performance metrics benefit from modifying the target loss function and incorporating AQI-level information. Overall, the improvement is still most pronounced for MAE and sMAPE. For the three windows, $(L_p, L_h) = (72, 24)$, $(96, 48)$, and $(120, 72)$, MAE is reduced by 1.1977%, 1.4150%, and 0.8887% on average, respectively,

for all the models compared with the MSE loss. The corresponding sMAPE is reduced by 0.9496%, 1.1689%, and 0.8307%, respectively. Moreover, the proposed IDW-ADAFormer again achieves the best MAE, RMSE, and sMAPE values in most cases. Overall, the relative ranking among different methods remains mostly consistent with those that were observed in Table 3.

Table 4. The forecasting performance of different models for PM_{10} time series under the AQI-aware loss.

Method	$(L_p, L_h) = (72, 24)$				$(L_p, L_h) = (96, 48)$				$(L_p, L_h) = (120, 72)$			
	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE	MAE	RMSE	AQI-Acc	sMAPE
IDW-ADAFormer	**2.1952**	**7.0003**	0.9997	**69.6904**	**2.3244**	**7.1119**	0.9998	**72.8543**	**2.3657**	**7.1879**	0.9997	**74.2665**
iTransformer	2.2462	7.0490	0.9994	70.7219	2.3318	7.1163	0.9999	73.3870	2.3868	7.1894	**1.0000**	74.6840
CrossFormer	3.9073	8.8589	0.9682	101.4268	5.1025	9.2510	0.9962	110.5981	5.0387	9.3783	0.9905	109.1889
DLinear	2.3305	7.1417	0.9994	73.3626	2.4451	7.1887	0.9998	77.0890	2.4851	7.2201	0.9999	78.4835
FreTS	2.3987	7.0416	**0.9999**	78.4097	2.5228	7.1237	**1.0000**	80.9736	2.6473	7.2130	0.9999	83.7681
TimeMixer	2.8094	7.2221	0.9989	85.7853	3.4622	7.6716	0.9987	95.2835	3.2548	7.6959	**1.0000**	89.4445
ETSformer	2.4629	7.0926	0.9238	86.1497	2.5697	7.1791	0.9132	90.2541	2.6811	7.3105	0.9132	92.5705
BigST	2.4314	7.0070	0.9995	75.5745	2.9108	7.1621	**1.0000**	88.2475	3.2645	7.3035	0.9999	95.0115
PatchSTG	3.4199	8.2723	0.9947	93.4438	3.8922	8.9001	0.9895	98.9754	4.1926	8.8173	0.9977	102.8080

Comparing the contributions of different attention components

Figure 3 and 4 illustrate the radar charts of the learned weights associated with different attention components when predicting the $PM_{2.5}$ and PM_{10} time series, respectively. The radar charts allow direct visual comparison of how the weighting patterns change under different optimization objectives. Both these figures consist of three subplots corresponding to the three windows considered, i.e., $(72, 24)$, $(96, 48)$, and $(120, 72)$, respectively.

There are two polygons shown in each subplot. The blue polygon represents the model parameters learned under the standard MSE loss. The red polygons are the model parameters obtained under the proposed multi-objective AQI-aware training. The five vertices of each polygon are labeled as α_{main}, α_{IDW}, α_{res}, β_{IDW}, and β_{res}, representing the learnable parameters in the attention mechanism as defined in (9) and (14), and as shown in Figure 1. Specifically, α_{main}, α_{IDW}, and α_{res} control the internal fusion of the CAK and the DCK output signals in the main, IDW, and residual branches, respectively. The parameters β_{IDW}, and β_{res}, on the other hand, govern the cross-branch fusion by modulating the contributions of the IDW and the residual branches when fusing their outputs into the main branch.

When predicting the $PM_{2.5}$ data under the MSE loss, α_{main}, α_{IDW} and α_{res} are close to 0.5, which indicates their balanced contributions from the CAK and DCK branches. The β values are smaller, suggesting a limited cross-branch influence. Provided that the training is done under the multi-objective AQI-aware loss, the α values are increased, so the model relies more on the CAK output than on the DCK output in each branch. At the same time, the β values decrease, so the influence of the auxiliary branches are decreased. Note also that the weight, $\beta_{\text{main}} = 1$, in Figure 1, which ensures that the main branch dominates the output of the ADA module.

For the PM_{10} time series, the overall patterns are similar to that of the $PM_{2.5}$ data. In particular, the α parameters are more balanced, whereas the β parameters generally decrease when assuming the AQI-aware loss. For both $PM_{2.5}$ and PM_{10} datasets, minor variations across different (L_p, L_h) settings can be observed.

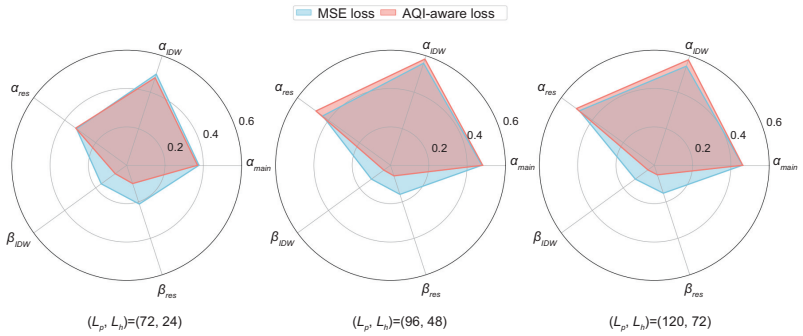


Figure 3. A radar chart of the fusion combining weights in predicting the $PM_{2.5}$ data.

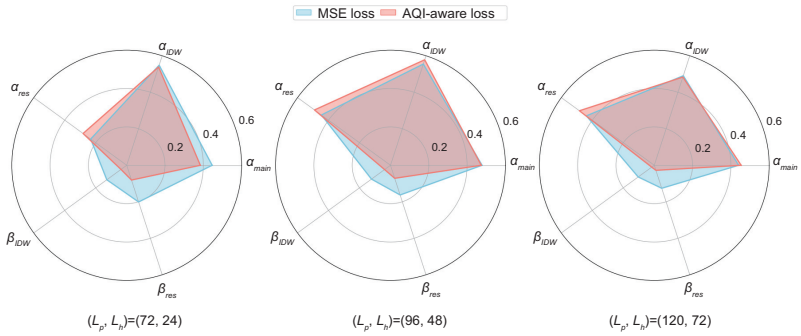


Figure 4. A radar chart of the fusion combining weights in predicting the PM_{10} data.

Discussion

The experimental results specifically consider predicting the $PM_{2.5}$ and PM_{10} time series. The proposed IDW-ADAFormer was found to generally outperform other competing architectures. The IDW-ADAFormer achieved the smallest MAE, RMSE, and sMAPE, while maintaining a high AQI-level classification accuracy. The IDW-ADAFormer is also more robust against the choice of the loss function.

More complex multi-objective AQI-aware loss function improves the forecasting performance for most models, which can be observed especially as a reduction in MAE and RMSE values. Since MAE and RMSE are particularly sensitive to large deviations, the reduction of their values indicates an improved robustness during unpredictable events having large variations. The joint regression-classification approach encourages learning more discriminative features, which improves both the forecasting accuracy and the model interpretability.

The radar chart analysis of relative contributions of the model branches shows that the CAK and the DCK contributions are generally balanced. The main branch in the ADA module provides the dominant predictive signal. The IDW and the residual branches act as minor adjustments, which is important for the improved forecasting accuracy. However, the influence of the IDW and the residual branches is reduced when the loss function becomes more complex.

The ablation studies revealed the contributions of individual attention components. The full model assumes $\mathbf{A}_{\text{main}}$, $\mathbf{A}_{\text{IDW}}$, and $\mathbf{A}_{\text{res}}$ signals, which leads to the most balanced and robust performance. The model variants excluding some of the attention components may have marginally smaller MAE, but then they underperform in terms of RMSE, sMAPE, and AQI-Acc. This indicates that the complete attention module is required to achieve the best and the most robust performance when predicting the extreme values in time series data. The extreme values have different effect on the evaluation metrics. For instance, RMSE disproportionately penalizes large errors, whereas sMAPE emphasizes relative deviations, and AQI-Acc measures the categorical consistency.

Despite the demonstrated advantages of the proposed IDW-ADAFormer, this architecture has several limitations. In particular, incorporating multiple attention branches with learnable parameters increases the training time, and it requires more training data. The increased computational complexity can be problematic when working with large datasets. The computing delays can negatively affect designing real-time forecasting applications. The effectiveness of the IDW and the residual branches depends on the ability of the attention module to extract good quality representative features, which is challenging, when the input data are very noisy. Lastly, it should be noted that these observations are mainly valid for the PM time series data. The effectiveness of the IDW-ADAFormer for other types of time series would requires performing additional experiments.

Conclusion

The IDW-ADAFormer was proposed to predict the multivariate time series generated by spatially separated sensors. The IDW-ADAFormer combines the adaptive k-nearest IDW decomposition, and a novel alignment-diversity attention. The IDW decomposition separates the spatial trends from the local temporal variations. The alignment-diversity attention module adaptively combines the outputs of the cross-sensor alignment kernel and the diversity-correlation kernel using learnable weights. The multi-objective AQI-aware loss function was shown it can improve the performance for medium-term forecasting of $PM_{2.5}$ and PM_{10} time series.

The experimental results showed that the proposed IDW-ADAFormer consistently outperforms other competing models in terms of MAE, RMSE, sMAPE, and AQI-level classification accuracy across different prediction horizons. The analysis of attention weights revealed that the CAK and the DCK contributions are balanced, and that the main attention branch provides the dominant prediction signal. The ablation studies confirmed that all components of the attention module are important for

achieving a good prediction performance.

The IDW-ADAFormer architecture can be further developed. Specifically, the multi-branch attention module dominates the overall computational complexity. Hence, future work may explore more efficient attention mechanisms to accelerate the kernel computations, such as sparse or adaptive attention to reduce unnecessary pairwise interactions, low-rank kernel approximations, and dynamic branch selection strategies to avoid full computation over all attention pathways while preserving the model expressiveness. It would also be useful to evaluate the prediction performance for other pollutants, such as NO_2 and O_3.

Funding Declaration

The research was funded by a research grant from Zhejiang University.

Author contributions

Y.Y.: Writing - original draft, Conceptualization, Investigation, Formal analysis, Software, Methodology, Data curation. P.L.: Writing - review & editing, Conceptualization, Investigations, Project administration, Supervision. Y. L.Y.: Writing - review & editing, Funding acquisition, Project administration, Supervision. G.: Resources, Project administration, Supervision. All authors reviewed the manuscript.

Additional information

The authors declare no competing interests.

Data availability

No new data were generated in the study. The PurpleAirSF public datasets were used to produce the numerical results, https://github.com/JingweiZuo/PurpleAirSF.

References

1. Shen, L., Wang, Y., Fan, X., Wei, Y. & Qiu, H. Variable-dynamic multivariate time-series forecasting for IoT systems. *IEEE Internet Things J.* **12**, 28479–28492 (2025). DOI 10.1109/JIOT.2025.3565737.
2. Joshi, D. *et al.* Fine particulate matter (PM2.5, PM10): A silent catalyst for chronic lung diseases in India; a comprehensive review. *Environ. Challenges* **20**, 101215 (2025). DOI 10.1016/j.envc.2025.101215.
3. Arca-Lafuente, S. *et al.* Effects of urban airborne particulate matter exposure on the human upper respiratory tract microbiome: a systematic review. *Respir. Res.* **26**, 118 (2025). DOI 10.1186/s12931-025-03179-9.
4. Jakhar, R., Styszko, K., Samek, L. & Szramowiat-Sala, K. Oxidative potential of PM1, PM2.5, and PM10 collected in car and tram tunnels to analyse their impact on public health. *Sci. Reports* **15**, 21773 (2025). DOI 10.1038/s41598-025-09037-4.
5. Ghosh, S. & Sinha, D. Indian perspective of PM2.5 attributed human health hazards during 2010–2025. *Air Qual. Atmosphere & Heal.* (2025). DOI 10.1007/s11869-025-01793-6.
6. Liang, Q. *et al.* Associations between ambient air pollutants and mortality from respiratory and cardiovascular diseases in the Shandong Peninsula urban agglomeration, China: A distributed lag non-linear model analysis. *Atmospheric Pollut. Res.* **16**, 102416 (2025). DOI 10.1016/j.apr.2025.102416.
7. Qu, S. *et al.* Pharmacotherapeutic strategies for fine particulate matter-induced lung and cardiovascular damage: Marketed drugs, traditional Chinese medicine, and biological agents. *Cardiovasc. Toxicol.* **25**, 666–691 (2025). DOI 10.1007/s12012-025-09985-3.
8. Dong, J., Zhang, Y. & Hu, J. Short-term air quality prediction based on EMD-transformer-BiLSTM. *Sci. Reports* **14**, 20513 (2024). DOI 10.1038/s41598-024-67626-1.
9. Li, S., Tang, B. & Deng, X. A hybrid method combining variational mode decomposition and deep neural networks for predicting PM2.5 concentration in China. *IEEE Access* **13**, 51956–51968 (2025). DOI 10.1109/ACCESS.2025.3552229.
10. Yue, X. *et al.* Novel hybrid data-driven modeling based on feature space reconstruction and multihead self-attention gated recurrent unit applied to PM2.5 concentrations prediction. *Sci. Reports* **15**, 17087 (2025). DOI 10.1038/s41598-025-00911-9.
11. Yin, Y. *et al.* PM-PCNet: A deep learning model for PM2.5 prediction integrating numerical forecast. In *Proceedings of CCDC*, 1–6 (2025). DOI 10.1109/CCDC65474.2025.11090403.
12. Lu, Y., Wang, J., Wang, D., Yoo, C. & Liu, H. Incorporating temporal multi-head self-attention convolutional networks and LightGBM for indoor air quality prediction. *Appl. Soft Comput.* **157**, 111569 (2024). DOI 10.1016/j.asoc.2024.111569.
13. Liu, Z., Ji, D. & Wang, L. PM2.5 concentration prediction based on EEMD-ALSTM. *Sci. Reports* **14**, 12636 (2024). DOI 10.1038/s41598-024-63620-9.

14. Wei, Q., Zhang, H., Yang, J., Niu, B. & Xu, Z. PM2.5 concentration prediction using a whale optimization algorithm based hybrid deep learning model in Beijing, China. *Environ. Pollut.* **371**, 125953 (2025). DOI 10.1016/j.envpol.2025.125953.

15. Liu, Y. *et al.* iTransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of ICLR* (2024).

16. Zhang, Y. & Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proceedings of ICLR* (2023).

17. Woo, G., Liu, C., Sahoo, D., Kumar, A. & Hoi, S. ETSformer: Exponential smoothing transformers for time-series forecasting (2022). ArXiv:2202.01381 [cs.LG].

18. Ma, Z. *et al.* Air pollutant prediction model based on transfer learning two-stage attention mechanism. *Sci. Reports* **14**, 7385 (2024). DOI 10.1038/s41598-024-57784-7.

19. Chen, P. *et al.* Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *Proceedings of ICLR* (2024).

20. Zhang, Y., Ma, L., Pal, S., Zhang, Y. & Coates, M. Multi-resolution time-series transformer for long-term forecasting. *CoRR* **abs/2311.04147** (2023).

21. Lin, M., Liu, P., Huang, C. & Lin, Y. The application of strategy based on LSTM for the short-term prediction of PM2.5 in city. *Sci. The Total. Environ.* **906**, 167892 (2024). DOI 10.1016/j.scitotenv.2023.167892.

22. Yang, J., Shi, L., Lee, J. & Ryu, I. Spatiotemporal prediction of particulate matter concentration based on traffic and meteorological data. *Transp. Res. Part D: Transp. Environ.* **127**, 104070 (2024). DOI 10.1016/j.trd.2024.104070.

23. Hong, W. Meteorological variability and predictive forecasting of atmospheric particulate pollution. *Sci. Reports* **14**, 14 (2024). DOI 10.1038/s41598-023-41906-8.

24. Wang, Y., He, H., Huang, H., Yang, J. & Peng, Z. High-resolution spatiotemporal prediction of PM2.5 concentration based on mobile monitoring and deep learning. *Environ. Pollut.* **364**, 125342 (2025). DOI 10.1016/j.envpol.2024.125342.

25. Fatima, N. *et al.* Time series AQI forecasting using Kalman-integrated Bi-GRU and Chi-square divergence optimization. *Sci. Reports* **15**, 29157 (2025). DOI 10.1038/s41598-025-12422-8.

26. Li, Y., Yang, J. & Qin, D. Accurate and interpretable PM2.5 prediction based on GC-AE-RegLSTM. *J. Comput. Sci.* **91**, 102687 (2025). DOI 10.1016/j.jocs.2025.102687.

27. Kurniawan, R. *et al.* Implementing the Haversine formula for detecting societal issues and delivering relevant preaching content. In *Proceedings of ICon-EEI*, 73–78 (2024). DOI 10.1109/IConEEI64414.2024.10748204.

28. Cipolla, R., Gal, Y. & Kendall, A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of CVPR*, 7482–7491 (2018). DOI 10.1109/CVPR.2018.00781.

29. U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division. Technical assistance document for the reporting of daily air quality – the air quality index (AQI). Tech. Rep., U.S. Environmental Protection Agency, Research Triangle Park, NC (2024).

30. Zuo, J., Li, W., Baldo, M. & Hacid, H. Unleashing realistic air quality forecasting: Introducing the ready-to-use PurpleAirSF dataset. In *Proceedings of AGIS* (2023). DOI 10.1145/3589132.3625575.

31. Zeng, A., Chen, M., Zhang, L. & Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of AAAI* (2023). DOI 10.1609/aaai.v37i9.26317.

32. Yi, K. *et al.* Frequency-domain MLPs are more effective learners in time series forecasting. In *Proceedings of NIPS* (Red Hook, NY, USA, 2023).

33. Wang, S. *et al.* TimeMixer: Decomposable multiscale mixing for time series forecasting. In *Proceedings of ICLR* (2024).

34. Han, J. *et al.* BigST: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proc. VLDB Endow.* **17**, 1081–1090 (2024). DOI 10.14778/3641204.3641217.

35. Fang, Y. *et al.* Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of KDD*, 307–317 (2025). DOI 10.1145/3690624.3709177.

36. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of ACM SIGKDD*, 2623–2631 (2019).

Figure legends

Figure 1. A block diagram of the proposed IDW-ADAFormer consists of the adaptive k-nearest IDW decomposition module and alignment-diversity attention module followed by iTransformer to generate predictions.

Figure 2. Block diagrams of (a) the canonical kernel attention, and (b) the proposed alignment-diversity attention.

Figure 3. A radar chart of the fusion combining weights in predicting the $PM_{2.5}$ data.

Figure 4. A radar chart of the fusion combining weights in predicting the PM_{10} data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf