atmosphere

MDPI

Article

Development of a Hybrid Attention Transformer for Daily $PM_{2.5}$ Predictions in Seoul

Hyun S. Kim *, Kyung M. Han, Jinhyeok Yu, Nara Youn and Taehoo Choi

School of Environment and Energy Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea; kmhan@gist.ac.kr (K.M.H.); jinhyeok.yu@gm.gist.ac.kr (J.Y.); narayoun@gm.gist.ac.kr (N.Y.); taehoo96@gm.gist.ac.kr (T.C.)

* Correspondence: hskim98@gist.ac.kr

Abstract: A hybrid attention transformer (HAT) was developed for accurate daily $PM_{2.5}$ predictions in Seoul. The performance of the HAT was evaluated through a comparative analysis of its predictions against ground-based observations and those from a three-dimensional chemical transport model (3-D CTM). The results demonstrated that the HAT outperformed the 3-D CTM, achieving a 4.60% higher index of agreement (IOA). Additionally, the HAT exhibited 22.09% fewer errors and 82.59% lower bias compared to the 3-D CTM. Diurnal variations in $PM_{2.5}$ predictions from both models were also analyzed to explore the characteristics of the proposed model further. The HAT predictions closely aligned with observed $PM_{2.5}$ throughout the day, whereas the 3-D CTM exhibited significant diurnal variability. The importance of the input features was evaluated using the permutation method, which revealed that the previous day's $PM_{2.5}$ was the most influential feature. The robustness of the HAT was further validated through a comparison with the long short-term memory (LSTM) model, which showed 18.50% lower errors and 95.91% smaller biases, even during El Niño events. These promising findings highlight the significant potential of the HAT as a cost-effective and highly accurate tool for air quality prediction.

Keywords: artificial neural network; hybrid attention transformer; daily $PM_{2.5}$ prediction

Academic Editors: Rana Muhammad Adnan, Ozgur Kisi and Mo Wang

Received: 3 December 2024
Revised: 27 December 2024
Accepted: 30 December 2024
Published: 1 January 2025

Citation: Kim, H.S.; Han, K.M.; Yu, J.; Youn, N.; Choi, T. Development of a Hybrid Attention Transformer for Daily $PM_{2.5}$ Predictions in Seoul. *Atmosphere* **2025**, *16*, 37. https://doi.org/10.3390/atmos16010037

1. Introduction

Fine particulate matter ($PM_{2.5}$), defined as particles with an aerodynamic diameter of less than 2.5 μm, has emerged as a significant public health concern due to its ability to cause serious adverse health effects [1,2]. $PM_{2.5}$ can infiltrate deep into the respiratory system and cross into the bloodstream, thereby increasing its potential to cause adverse effects on human health [3,4]. Numerous epidemiological studies have demonstrated a strong correlation between elevated $PM_{2.5}$ and the prevalence of respiratory and cardiovascular diseases [4–7]. Consequently, heightened $PM_{2.5}$ concentrations have drawn increasing attention from both researchers and policymakers due to the associated health risks and social implications. Since 2014, high levels of $PM_{2.5}$ have become a critical social issue in South Korea.

To minimize the public health risks associated with high $PM_{2.5}$ levels, the National Institute of Environmental Research (NIER) of South Korea has been conducting operational air quality forecasts using a three-dimensional chemistry model (3-D CTM) since 2014. However, despite the widespread application of CTMs, these models exhibit inherent limitations in accurately predicting high $PM_{2.5}$. The discrepancies in 3-D CTM predictions arise from several sources of uncertainty, including emissions, meteorological fields, initial

and boundary conditions, and physicochemical mechanisms. While significant efforts have been made to enhance the predictive performance of 3-D CTMs [8–11], the exact timeline for achieving reliable and sufficiently accurate air quality forecasts remains unclear.

Accurate $PM_{2.5}$ prediction is crucial for air quality management and public health. Given the limitations of low predictive accuracy in traditional CTMs, researchers have turned to alternative methods capable of handling large datasets and complex interactions among input variables. In recent years, artificial intelligence (AI) approaches have gained significant attention in the field of air quality prediction due to their superior performance and computational efficiency compared to traditional CTMs. Early AI applications primarily relied on simple machine learning (ML) algorithms, such as multilayer perceptron (MLP) and the autoregressive integrated moving average (ARIMA), to predict ambient $PM_{2.5}$ levels [12–14]. However, these approaches have exhibited notable limitations: while MLP is capable of capturing nonlinear relationships between input features, it struggles with temporal dependencies and the vanishing gradient problem. Additionally, ARIMA, which assumes data stationarity, inadequately represents the nonlinear dynamic nature of atmospheric variables. Hence, more advanced techniques have been sought to overcome these drawbacks and better exploit the hidden patterns in large-scale air quality data.

To address these shortcomings, recent advancements in AI have introduced more sophisticated deep neural network (DNN) architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and hybrid RNN–CNN models, for $PM_{2.5}$ predictions [15–20]. In particular, RNNs have been proven to be well-suited for time-series predictions due to their ability to capture temporal dependencies through internal memory cells [15–17]. Among RNNs, long short-term memory (LSTM) networks have gained widespread adoption due to their effectiveness in addressing the issues of vanishing and exploding gradients, which often hinder long-term learning in standard RNNs [15,16,21,22]. However, despite these strengths, RNN-based methods alone may not be sufficient to capture spatial correlations, especially in regions where $PM_{2.5}$ levels vary significantly, such as East Asia.

In addition, CNNs have demonstrated significant effectiveness in $PM_{2.5}$ predictions by capturing spatial dependencies within multidimensional data. By extracting spatial features from air quality data across different locations, CNNs are able to identify patterns associated with air pollutant distribution and movement [23]. Moreover, combining CNNs with LSTM in hybrid models enables simultaneous modeling of both spatial and temporal dependencies, thereby improving the accuracy of $PM_{2.5}$ predictions. This hybrid approach capitalizes on CNNs' strength in capturing spatial patterns and LSTMs' capability in modeling temporal dynamics, making it especially suited for the complex, multi-dimensional nature of air quality data [19,20,24,25].

Despite these advancements, each of these DNN approaches present inherent limitations. While LSTM networks effectively capture temporal dependencies, they face challenges related to long-term memory retention and computational efficiency. CNNs, which excel at spatial feature extraction, are less effective at modeling sequential dynamics. Although the CNN–LSTM hybrid approach addresses both spatial and temporal dependencies, it is computationally intensive. Furthermore, these architectures remain susceptible to the issue of vanishing gradients [26]. Consequently, there is a critical need for the development of novel architectures that preserve the strengths of deep learning while systematically addressing these shortcomings.

Recently, transformer architecture has emerged as a promising solution to overcome the limitations of modeling complex spatiotemporal dependencies. Unlike RNNs, which process data sequentially, transformers leverage self-attention mechanisms that allow for the simultaneous capture of dependencies across all time steps, thereby overcoming limita-

tions associated with long-term memory retention and computational inefficiency [27]. The self-attention mechanism dynamically assigns varying degrees of importance to different input features, enabling the model to focus on crucial information and establish relationships across distinct instances within a sequence. Additionally, the incorporation of residual connections within the transformer architecture improves gradient flow during backpropagation, facilitating the training of deep networks and enhancing convergence rates [28]. By integrating self-attention with residual connections, the transformer architecture captures both temporal and spatial dynamics in a computationally efficient manner.

These advantages have contributed to the growing use of transformer-based algorithms for $PM_{2.5}$ prediction in recent research [29,30]. Several studies have compared the performance of transformer-based models with traditional algorithm-based approaches for $PM_{2.5}$ prediction [31–35] (see Table 1). According to these studies, transformers demonstrated approximately 47% improvement in prediction accuracy, as measured by root mean squared error (RMSE). Such promising findings highlight the potential of transformer models in providing more reliable predictions, and they serve as a strong motivation for further exploration and refinement in this area.

Table 1. Performance comparison between transformer-based models and traditional AI algorithms for $PM_{2.5}$ predictions.

References	Period	Study Area	Algorithms [1]	Metrics [2]	Accuracy
Cui et al., 2023 [31]	2016–2017	Beijing, China	CNN + LSTM + Attention Transformer	RMSE	6.85 3.28
Wang et al., 2023 [32]	2016–2022	Beijing, China	SVM RF AdaBoost LSTM GRU Transformer	RMSE	19.67 23.00 21.62 20.79 18.05 11.11
Rai et al., 2023 [33]	2016–2020	New Delhi, India	RNN LSTM BiLSTM Transformer	MAPE	47.01 27.01 13.40 7.69
Dai et al., 2024 [34]	2021–2023	Zhengzhou, China	MLP LSTM GNN Transformer	RMSE	8.81–11.14 7.72–11.01 8.35–12.17 3.51–11.00
Zou et al., 2024 [35]	2020–2022	Yangtze River Delta, China	MLP LSTM Transformer	RMSE	13.26 12.33 7.27

[1] SVM stands for support vector machine; RF stands for random forest; AdaBoost stands for adaptive boosting; GRU stands for gated recurrent unit; BiLSTM stands for bidirectional long short-term memory; GNN stands for graph neural network. [2] The units for RMSE and MAPE (mean absolute percentage error) are $\mu g/m^3$ and %, respectively.

The aim of this study is to develop and optimize a hybrid attention transformer (HAT) model to achieve more accurate daily $PM_{2.5}$ predictions in Seoul. This model integrates the benefits of transformer architecture with a hybrid attention mechanism to effectively capture both temporal and spatial dependencies in complex atmospheric datasets. The detailed methodology and configuration of the HAT model are provided in Section 2.

2. Model Development

2.1. Input Features

We constructed a 5.3-year dataset to optimize the HAT model (January 2019 to December 2023) and to make predictions (January 2024 to April 2024). The input features were sourced from publicly accessible observational datasets, including the Korea Meteorological Administration's (KMA) Automated Surface Observing System (ASOS) and the NIER's AIR KOREA network. The datasets were acquired from their official archives (https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36, accessed on 20 November 2024; https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123, accessed on 20 November 2024), respectively. By utilizing these public datasets, this study aims to achieve cost-effectiveness, reduce technical barriers in air quality prediction, and facilitate the widespread adoption of the developed framework by researchers, policymakers, and environmental agencies.

Figure 1 illustrates the locations of the ASOS and AIR KOREA monitoring stations in Seoul. As shown, there is one meteorological monitoring station and 25 air quality monitoring stations, all of which correspond to urban sites and provide hourly information. The ASOS and AIR KOREA networks were utilized in this study as they provide essential and reliable datasets, providing comprehensive coverage of meteorological and air quality conditions in Seoul, which are critical for accurate $PM_{2.5}$ prediction. A summary of the input features and their statistical characteristics used in the HAT-based $PM_{2.5}$ prediction is provided in Table 2. Consistent with this study, previous research has demonstrated that integrating input features from both observation-based meteorological data and air quality information can achieve cost-effectiveness and accuracy in $PM_{2.5}$ prediction using DNNs [16–20,31–35].

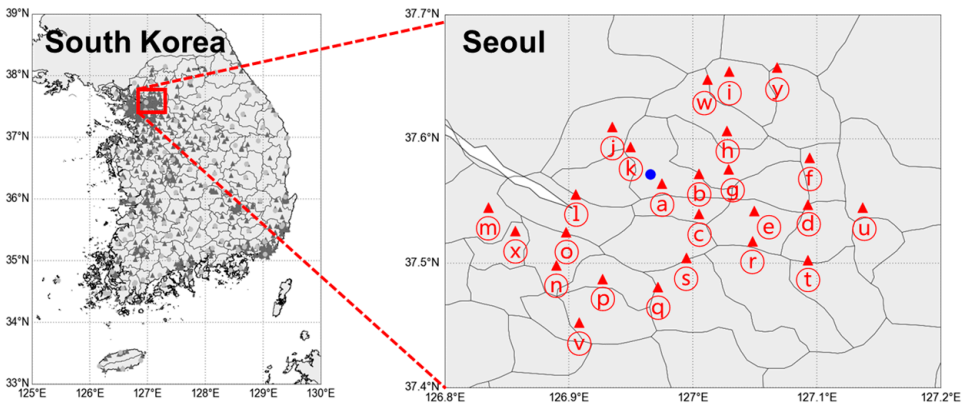

Figure 1. Locations of the KMA ASOS and NIER AIR KOREA ground monitoring stations in Seoul. The blue circle and red triangles represent ASOS and AIR KOREA sites, respectively. Circled alphabetic characters denote unique identifiers assigned to each AIR KOREA monitoring station.

The configured input features are closely linked to ambient $PM_{2.5}$, providing valuable insights into atmospheric conditions. Among the meteorological features, wind speed plays a crucial role in pollutant dispersion and advection, influencing turbulent mixing and the horizontal transport of particulate matter. Wind direction, which indicates the origin of air masses, assists in tracking pollutant movement across regions. Precipitation, a key wet scavenging mechanism, reduces $PM_{2.5}$ levels by removing airborne pollutants. Relative humidity affects the hygroscopic growth of ambient aerosols, influencing their size and chemical composition. Temperature and pressure provide insights into seasonal variations and weather system dynamics. Air quality features, such as concentrations of SO_2, CO, and NO_2, reflect the intensity of anthropogenic emissions and contribute to the formation of particulate matter, which correlates with $PM_{2.5}$ levels. O_3 levels reveal the strength of atmo-

spheric oxidation processes that form secondary particulate matter, including sulfates and nitrates. Additionally, observed $PM_{2.5}$ and PM_{10} provide important context for predicting future $PM_{2.5}$, highlighting the persistence of pollutants in the atmosphere. Incorporating the previous day's PM levels allows models to capture temporal dynamics, improving predictive accuracy. By considering both meteorological and air quality features, a more precise understanding of their interdependencies is achieved, enhancing the prediction of $PM_{2.5}$.

Table 2. Summary of input features and their statistical characteristics.

Feature Category	Features	Units	Temporal Resolution	Minimum	Maximum	Mean ± SD *
Meteorological feature	Temperature	°C	1 h	−18.50	36.70	13.61 ± 10.79
	Precipitation	mm	1 h	0.00	64.70	0.61 ± 1.29
	Wind speed	m/s	1 h	0.00	9.10	2.26 ± 1.14
	Wind direction	°	1 h	0.00	360.00	189.75 ± 109.30
	Relative humidity	%	1 h	10.00	100.00	63.34 ± 19.63
	Pressure	hPa	1 h	975.20	1028.00	1005.98 ± 8.25
Atmospheric environmental feature	SO_2	ppmv	1 h	0.00	0.02	0.00 ± 0.00
	CO	ppmv	1 h	0.02	2.70	0.47 ± 0.21
	O_3	ppmv	1 h	0.00	0.23	0.03 ± 0.02
	NO_2	ppmv	1 h	0.00	0.12	0.02 ± 0.01
	PM_{10}	$\mu g/m^3$	1 h	0.00	1024.00	37.08 ± 31.41
	$PM_{2.5}$	$\mu g/m^3$	1 h	0.00	237.00	20.62 ± 16.61

* SD stands for standard deviation.

To mitigate the risk of overestimating the influence of features with relatively higher values, we normalized the input features to a range of −1 to 1 using the following equation:

$$x_{normal,\,i} = 2 \times \left(\frac{x_i - x_{min,i}}{x_{max,i} - x_{min,i}} \right) - 1 \quad (1)$$

here $x_{normal,\,i}$ represents the normalized value of feature i, x_i denotes the original value of feature i, and $x_{max,i}$ and $x_{min,i}$ refer to the maximum and minimum values of feature i, respectively.

2.2. Hybrid Attention Transformer

The overall structure of the proposed HAT model is illustrated in Figure 2. As shown, the HAT consists of three main components: (i) embedding, (ii) encoder, and (iii) decoder. By systematically integrating these components, the model transforms raw input features into progressively more informative representations, thereby enabling accurate $PM_{2.5}$ predictions.

The embedding component integrates both shallow feature extraction and positional encoding (PE) processes. For daily predictions, the input features were configured with dimensions of 25 × 24 × 12, corresponding to the number of AIR KOREA monitoring stations, the hours of the day, and the number of features, respectively. Shallow feature extraction was achieved by utilizing a single dense layer (fully connected layer) with 64 hidden nodes, which efficiently transforms the raw input features into intermediate features. This initial transformation step is crucial for capturing and refining the essential signals in the dataset, thereby laying a strong foundation for subsequent predictive tasks.

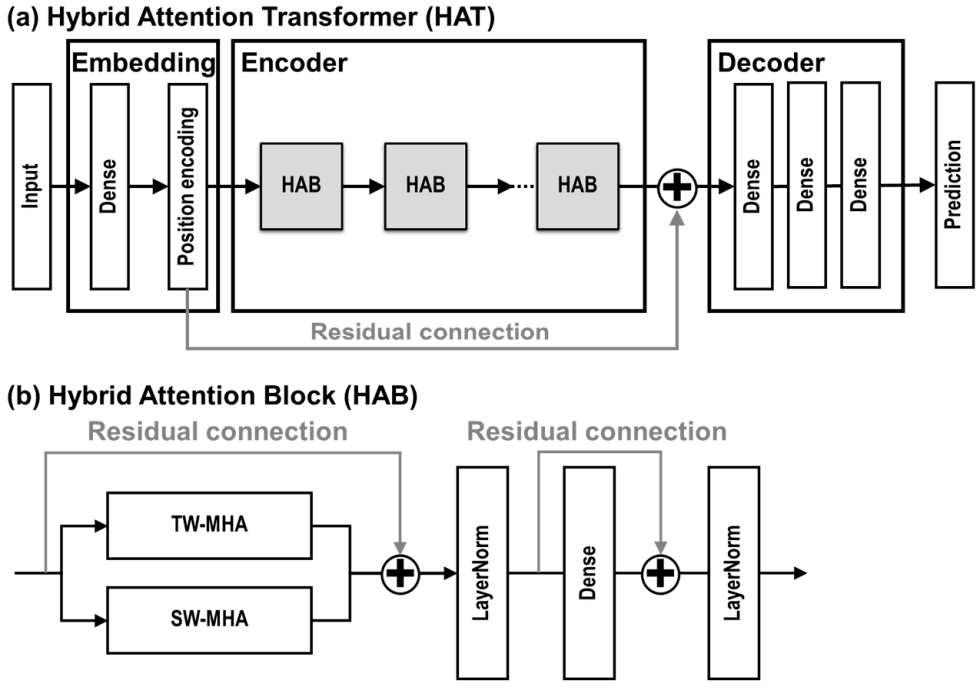

Figure 2. The overall architecture of the $PM_{2.5}$ prediction model: (**a**) HAT and (**b**) HAB.

As illustrated in Figure 2, the HAT model does not incorporate any recurrent or convolutional operations. Therefore, it is essential to integrate information regarding the relative and/or absolute positions of tokens within the intermediate sequence to achieve a more precise representation and understanding of the sequence structure. To address this need, a positional encoding block was utilized to integrate positional information directly into the shallow features. This approach enables the transformer to effectively capture token order and dependencies, even without traditional sequential processing. The positional information of the intermediate sequence is estimated using the sinusoidal method, as described by the following equations [27]:

$$PE_{(POS,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

$$PE_{(POS,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{3}$$

where pos represents the position index in the sequence, i denotes the dimension in the embedding, and d_{model} represents the total dimension of the embedding space. As shown in the above equations, the frequencies decrease along the vector dimension, forming a geometric progression from 2π to $10000 \cdot 2\pi$ in terms of wavelengths. This encoding approach enables the transformer-based predictive models to incorporate both frequency and positional information into their estimations effectively.

The position-encoded shallow features are fed into the encoder and linked to the encoder's final output via a residual (or skip) connection. The integration of residual connections significantly enhances the performance of the transformer architecture by mitigating the vanishing gradient problem, a common challenge in optimizing DNNs. This architectural choice promotes more efficient gradient propagation throughout the network, enabling the effective training of deeper and more sophisticated structures. Furthermore, residual connections preserve essential information across layers, ensuring that crucial features are retained as they propagate through the network.

The second component of the HAT is responsible for extracting deep features, referred to as latent representations, from the shallow features. To achieve this, the encoder was designed with five hybrid attention blocks (HABs), determined through a comprehensive sensitivity analysis that evaluated validation metrics while varying the number of HABs. This analysis identified five HABs as achieving the highest predictive accuracy. The details of each HAB are illustrated in Figure 2b. In this study, we implemented a hybrid attention mechanism consisting of two distinct modules: time-wise multi-head self-attention (TW-MHA) and station-wise multi-head self-attention (SW-MHA). TW-MHA and SW-MHA compute attention scores that represent the importance of temporal and spatial features, respectively. The complete computation process of the HAB is as follows:

$$X_M = \alpha TW\text{-}MHA(X) + \beta SW\text{-}MHA(X) + X \tag{4}$$

$$X_N = LayerNorm(X_M) \tag{5}$$

$$X_O = Dense(X_N) + X_N \tag{6}$$

$$Y = LayerNorm(X_O) \tag{7}$$

where X represents the features extracted from the previous stage; X_M, X_N, and X_O denote the intermediate features within the HAB; Y is the output of HAB; and $TW\text{-}MHA$ and $SW\text{-}MHA$ refer to time-wise and station-wise attention operations, respectively. $LayerNorm$ refers to the layer normalization operation, which stabilizes the learning process by normalizing the inputs across the features for each training sample, thereby reducing internal covariate shifts. $Dense$ refers to the fully connected layer, which transforms the normalized features into a new representation. The learnable weights α and β, initialized at 0.5, are fine-tuned during training to balance the influence of each attention module dynamically. This flexible approach enables the model to allocate attention adaptively and contextually, ensuring optimal performance across diverse feature interactions.

Both attention modules utilized a multi-head self-attention mechanism to compute attention scores, enabling the model to focus on multiple aspects of the data simultaneously. This ensures that no single feature or position disproportionately influences the learning process. The attention scores were calculated using the following equations:

$$MultiHead(Q,\ K, V) = Concat(\text{head}_1,\ \ldots,\ \text{head}_h)W^O \tag{8}$$

$$\text{head}_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{9}$$

$$Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) = SoftMax\left(\frac{QW_i^Q\left(KW_i^K\right)^T}{\sqrt{d_{k_i}}}\right)VW_i^V \tag{10}$$

here Q, K, and V are the query, key, and value matrices; W_i^Q, W_i^K, and W_i^V are learnable linear transformation matrices for the i-th head; d_{k_i} is the dimensionality of the key vectors for the i-th attention head; and W^O is the output weight matrix that linearly transforms the concatenated outputs of all attention heads into the desired dimensionality. $Concat$ refers to the concatenating the outputs of each attention head along the feature dimension. $SoftMax$ represents the softmax function, which normalizes the attention scores and converts them into probabilities that represent the attention distribution across all keys. In this study, the number of attention heads was set to 8, determined through a comprehensive sensitivity analysis, which revealed that eight attention heads achieved the highest accuracy.

The original input features had a dimension of $25 \times 24 \times 12$, where 25 represents the number of AIR KOREA stations, 24 denotes the number of hours in a day, and 12 corresponds to the number of input features. Expanding from 12 to 64 features en-

ables the network to capture a broader range of patterns, thereby enhancing its capacity to detect subtle variances in the dataset. After the embedding process, the shallow, intermediate features had a dimension of $25 \times 24 \times 64$, where 64 represents the number of extracted shallow features. Sensitivity tests involving different numbers of hidden nodes identified 64 as the optimal choice based on the lowest validation mean squared error (MSE), ensuring a balance between model capacity and computational efficiency. For SW-MHA, the shallow features were further transformed into a $24 \times 25 \times 64$ format, representing time, station, and intermediate features. This transformation facilitates a detailed analysis of spatial dependencies by examining the relationships between observation points, thereby enhancing our understanding of spatial dependencies within the data. Consequently, this approach enables the HAB to focus on different aspects of the input features, leading to a more nuanced understanding of both temporal and spatial dependencies.

To clearly demonstrate the advantages of the HAB, we evaluated $PM_{2.5}$ prediction performance using both individual attention modules and the HAB. The results of this comparative analysis are summarized in Table 3. The HAB achieved a substantially lower prediction error of 12.82 $\mu g/m^3$ and bias of 0.26 $\mu g/m^3$, whereas the individual attention modules produced errors ranging from 13.90 $\mu g/m^3$ to 15.02 $\mu g/m^3$ and biases from 3.89 $\mu g/m^3$ to 3.97 $\mu g/m^3$. These results highlight the HAB's ability to more effectively capture both temporal and spatial dependencies in the atmospheric dataset. By leveraging the hybrid attention structure, the HAB selectively focuses on the most pertinent features across various scales, enhancing both interpretability and robustness. This improved representational capability not only facilitates more accurate $PM_{2.5}$ predictions but also provides deeper insights into the underlying latent feature relationships.

Table 3. Comparison of $PM_{2.5}$ prediction performance across different attention methods.

Performance Metrics	SW-MHA	TW-MHA	HAB
IOA	0.74	0.73	0.75
R	0.58	0.55	0.59
RMSE	13.90	15.02	12.82
MB	3.89	3.97	0.26
PBIAS	17.09	17.57	1.13

IOA and R are dimensionless metrics; the units for RMSE and MB are $\mu g/m^3$; the unit for PBIAS is %.

The latent features extracted from the encoder were then fed into the decoder component, which is responsible for identifying the nonlinear relationships between the latent information and the true values. To optimize the decoder architecture for this task, sensitivity tests were conducted by systematically varying layer sizes and activation functions to minimize validation mean squared error (MSE). Based on these tests, the decoder was constructed with three consecutive dense layers, consisting of 128, 64, and 1 unit, respectively. To enable this nonlinear transformation, a leaky rectified linear unit (leaky ReLU) activation function was employed between each pair of dense layers. Leaky ReLU was chosen for its ability to mitigate the vanishing gradient problem by allowing a small, nonzero gradient when a unit is inactive, thereby addressing the dying ReLU problem. This property facilitates more efficient optimization, particularly in complex and deeper networks. The mathematical expression for the leaky ReLU is:

$$f(x) = \begin{cases} \alpha x & when\ x < 0 \\ x & when\ x \geq 0 \end{cases} \tag{11}$$

where x is the input, and α is the slope for negative values. In this study, α was set to 0.01.

2.3. Optimization

Model optimization seeks to determine the optimal combination of adaptive learnable parameters, including weight and bias matrices, through an iterative refinement process. The architecture of the proposed HAT model, encompassing both structural and optimization-related hyperparameters, was meticulously designed through extensive sensitivity analyses to ensure its suitability for daily $PM_{2.5}$ prediction tasks. These analyses helped identify an optimal set of hyperparameters that balanced predictive accuracy and computational feasibility. Key structural hyperparameters—such as the number of hidden nodes, the number of HABs, the number of attention heads, and other essential components—were systematically evaluated and fine-tuned to enhance the model's predictive performance, as detailed in Section 2.2. Additionally, optimization-related hyperparameters—including batch size, number of epochs, learning rate, and weight initialization method—were systematically evaluated to ensure computational efficiency and training stability. During this process, various configurations were tested and validated, while training and validation losses were closely monitored to mitigate risks of overfitting and underfitting. A batch size of 24 was selected to maintain computational efficiency and gradient stability. The learning rate of 0.5×10^{-4} was chosen to balance convergence speed and stability. The number of epochs was limited to 100, incorporating early stopping to prevent overfitting. He initialization was employed due to its superior gradient stability, supporting efficient and robust training [36]. Hence, each of these choices was guided by a comprehensive assessment of the model's learning dynamics and error profiles.

In the context of DNN optimization, the loss function—also known as the cost function or objective function—plays a vital role. This function quantifies the discrepancy between the model's predicted and actual values, providing a critical metric to guide the optimization process. By minimizing this error, the loss function facilitates the adjustment of the model's adaptive parameters, thereby improving the model's accuracy.

For the proposed transformer, which was designed for regression tasks, the MSE was selected as the loss function. MSE is particularly suitable for regression problems as it quantifies the average squared difference between predicted and actual values. This characteristic allows MSE to assign greater weight to larger prediction errors, making it especially useful for identifying and reducing significant discrepancies between predictions and actual values. Additionally, MSE is sensitive to outliers, encouraging the model to minimize larger errors that could otherwise have an outsized impact on overall performance. This characteristic is critical in $PM_{2.5}$ prediction, where extreme pollutant concentrations pose heightened public health risks. MSE is expressed as:

$$J_{MSE}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - h_\theta(x_i))^2 \tag{12}$$

here y_i denotes the true value, and $h_\theta(x_i)$ represents the prediction generated by the HAT model given the input x_i.

To effectively determine the optimal combination of the model's learnable parameters, the adaptive moment estimation (Adam) optimizer was employed [37]. Adam is a widely adopted optimization algorithm due to its adaptive learning rate and computational efficiency. It combines the benefits of both momentum and the root mean square propagation (RMSProp) algorithms by adjusting the learning rate for each parameter individually based on its past gradients. This flexibility enables the model to navigate the loss surface effectively, adapting to unique gradient patterns of different parameters. This adaptive behavior is particularly advantageous in training deep neural networks, such as the proposed HAT model, where parameter gradients can vary significantly throughout the training process.

By leveraging the Adam optimizer to minimize the MSE loss function, the model effectively reduces the discrepancy between its predictions and the actual values, enhancing its overall performance and generalizability. Ultimately, this helps the DNN converge faster and more reliably, mitigating issues like overshooting or slow progress.

In optimizing the HAT model, a 5-year dataset spanning from January 2019 to December 2023 was used. This dataset was divided into two subsets: 80% allocated for training the model and 20% reserved for validating its performance. This split provides sufficient data for the model to learn patterns while retaining a meaningful portion for unbiased evaluation. Throughout the optimization process, the variations in MSE loss for both datasets were assessed continuously to ensure ongoing enhancement of the model's effectiveness. During the early stages of model training, both the training and validation MSEs decrease as the model's weights and biases are iteratively adjusted. Since the training dataset is solely involved in updating the learnable parameters, the training MSE consistently decreases with each training iteration. However, the validation MSE follows a different pattern. Initially, it decreases alongside the training MSE, but after a certain point, it begins to either stagnate or increase. This phenomenon is commonly observed in the optimization of DNNs, where the model initially improves but eventually faces diminishing returns or overfitting. The iterative process of parameter updates continues until this plateau or inflection point for the validation MSE is reached. Beyond this point, additional training risks degrading generalization performance, necessitating model finalization to avoid overfitting.

2.4. CTM-Based $PM_{2.5}$ Prediction

A comparative analysis was conducted to evaluate the performance of $PM_{2.5}$ predictions produced by the HAT model against those from a 3-D CTM. Specifically, the community multiscale air quality (CMAQ) model, version 5.2.1, was employed to produce CTM-based $PM_{2.5}$ predictions, serving as a benchmark for assessing the accuracy of the HAT model. This analysis offered valuable insights into the effectiveness of the HAT approach relative to traditional CTM-based methods for predicting daily $PM_{2.5}$. The meteorological inputs for the CTM simulations were derived from the weather research and forecasting (WRF) model, version 4.1.5. Figure 3 illustrates the domain for the CTM-based $PM_{2.5}$ predictions, which encompasses the Northeast Asia region. The CTM predictions were configured with a horizontal resolution of 15 × 15 km². Anthropogenic emission data were sourced from the ASIA-AQ v3.0 emission inventory, which was developed to support the Airborne and Satellite Investigation of Asian Air Quality (ASIA-AQ) field campaign. Biogenic emissions were derived from the model of emissions of gases and aerosols from nature (MEGAN) v2.1 [38].

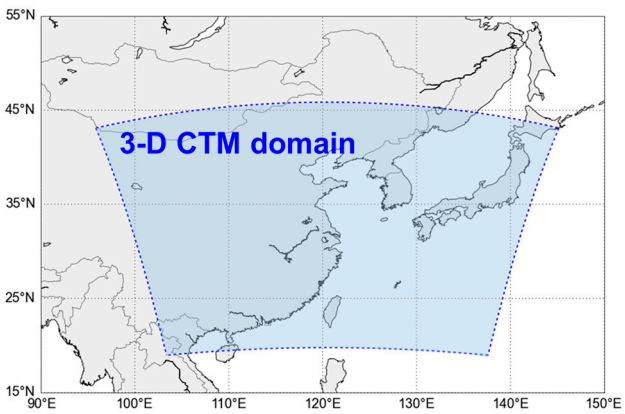

Figure 3. Spatial coverage of the 3-D CTM-based $PM_{2.5}$ predictions.

3. Results

3.1. Performance Evaluation

Daily $PM_{2.5}$ predictions were performed over a 4-month period, spanning from January 2024 to April 2024. The effectiveness of the proposed transformer was assessed by comparing the $PM_{2.5}$ predictions from the HAT with ground observations and estimates from the 3-D CTM. To rigorously evaluate the accuracy of the HAT-based $PM_{2.5}$ predictions, five statistical metrics were employed: index of agreement (IOA), Pearson correlation coefficient (R), RMSE, mean bias (MB), and percent bias (PBIAS) [39–41]. These metrics offer complementary insights into model performance, enabling a comprehensive evaluation of its strengths and limitations. IOA measures the normalized concordance between observed and predicted values, capturing both systematic and random deviations. R quantifies the strength and direction of the linear relationship between observations and predictions, indicating their alignment with actual trends. RMSE evaluates the magnitude of prediction errors, reflecting model precision. MB identifies systematic over- or underestimations, highlighting consistent trends in model behavior. PBIAS expresses mean bias relative to observed values, providing a dimensionless metric particularly useful for comparative studies. The mathematical formulations for these metrics are presented as follows:

$$IOA = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \tag{13}$$

$$R = \frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2 \sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}} \tag{14}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{15}$$

$$MB = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i) \tag{16}$$

$$PBIAS = 100 \times \frac{\sum_{i=1}^{N}(P_i - O_i)}{\sum_{i=1}^{N} O_i} \tag{17}$$

here P_i and O_i represent the model predicted and observed $PM_{2.5}$, and $\overline{P}$ and $\overline{O}$ denote the mean values of P_i and O_i.

Figure 4 and Table 4 provide detailed evaluations of $PM_{2.5}$ prediction performance across 25 monitoring stations. In the figure, black open circles with a dashed line represent the observed $PM_{2.5}$, and the blue and red lines correspond to the $PM_{2.5}$ predictions from the 3-D CTM and the HAT, respectively. Among the statistical metrics, IOA serves as a robust, dimensionless measure of agreement between predicted and observed values. IOA is particularly valuable because it simultaneously considers correlations, errors, and biases, offering a comprehensive assessment of model performance. Higher IOA values indicate more reliable predictions, underscoring its utility in comparing model outputs from various perspectives.

The comparison demonstrated that the HAT-based $PM_{2.5}$ predictions consistently outperformed the 3-D CTM-based predictions. The IOA values for the HAT model ranged from 0.71 to 0.77, while the 3-D CTM predictions exhibited IOA values between 0.69 and 0.74. On average, the HAT achieved higher IOA values by 4.60% compared to the 3-D CTM, indicating statistically stronger alignment with the observed $PM_{2.5}$. The highest IOA values for the HAT model were recorded at stations b, g, and l, where IOA reached 0.77, supported by R values of 0.62 to 0.63. In contrast, station j demonstrated the lowest IOA

value of 0.71 for the HAT model; however, this still exceeded the 3-D CTM's performance at the same station. Even at stations f, i, and u, where the 3-D CTM exhibited relatively better accuracy, the HAT model consistently outperformed it, achieving IOA values higher by 0.01 to 0.02. These findings highlight the HAT model's capability to achieve superior $PM_{2.5}$ prediction performance across diverse temporal and spatial conditions.

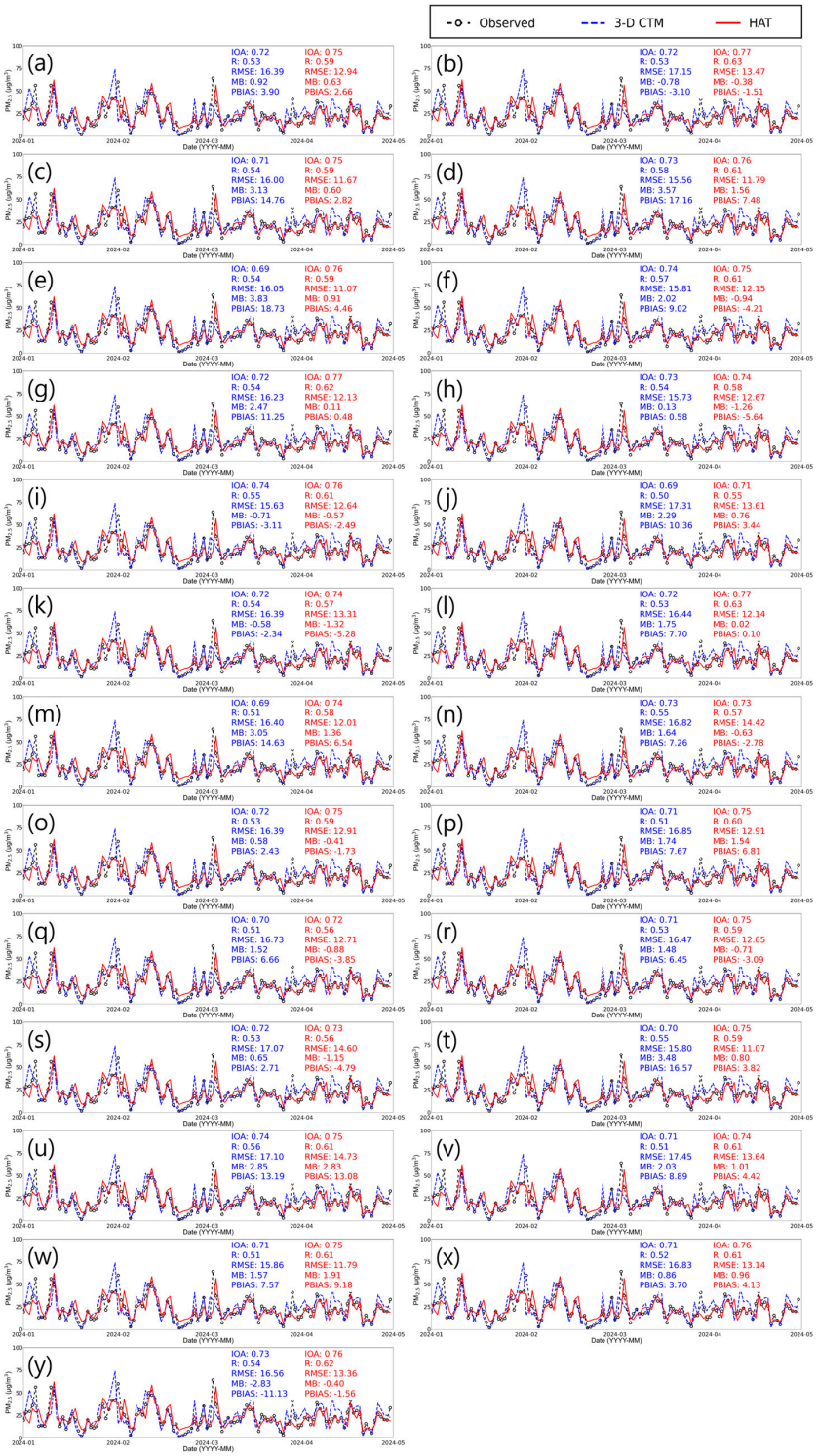

Figure 4. Comparisons of observed and predicted $PM_{2.5}$ in Seoul over the 4-month prediction period. Black open circles with dashed lines indicate observed $PM_{2.5}$, and blue and red lines represent the 3-D CTM- and HAT-predicted $PM_{2.5}$, respectively. Panels (**a**–**y**) correspond to individual AIR KOREA stations, as referenced in Figure 2.

Table 4. Monitoring station-wise evaluation of $PM_{2.5}$ predictions from the HAT and 3-D CTM *.

Metrics	Station a		Station b		Station c		Station d		Station e	
	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT
IOA	0.72	0.75	0.72	0.77	0.71	0.75	0.73	0.76	0.69	0.76
R	0.53	0.59	0.53	0.63	0.54	0.59	0.58	0.61	0.54	0.59
RMSE	16.39	12.94	17.15	13.47	16.00	11.67	15.56	11.79	16.05	11.07
MB	0.92	0.63	−0.78	−0.38	3.13	0.60	3.57	1.56	3.83	0.91
PBIAS	3.90	2.66	−3.10	−1.51	14.76	2.82	17.16	7.48	18.73	4.46
Metrics	**Station f**		**Station g**		**Station h**		**Station i**		**Station j**	
	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT
IOA	0.74	0.75	0.72	0.77	0.73	0.74	0.74	0.76	0.69	0.71
R	0.57	0.61	0.54	0.62	0.54	0.58	0.55	0.61	0.50	0.55
RMSE	15.81	12.15	16.23	12.13	15.73	12.67	15.63	12.64	17.31	13.61
MB	2.02	−0.94	2.47	0.11	0.13	−1.26	−0.71	−0.57	2.29	0.76
PBIAS	9.02	−4.21	11.25	0.48	0.58	−5.64	−3.11	−2.49	10.36	3.44
Metrics	**Station k**		**Station l**		**Station m**		**Station n**		**Station o**	
	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT
IOA	0.72	0.74	0.72	0.77	0.69	0.74	0.73	0.73	0.72	0.75
R	0.54	0.57	0.53	0.63	0.51	0.58	0.55	0.57	0.53	0.59
RMSE	16.39	13.31	16.44	12.14	16.40	12.01	16.82	14.42	16.39	12.91
MB	−0.58	−1.32	1.75	0.02	3.05	1.36	1.64	−0.63	0.58	−0.41
PBIAS	−2.34	−5.28	7.70	0.10	14.63	6.54	7.26	−2.78	2.43	−1.73
Metrics	**Station p**		**Station q**		**Station r**		**Station s**		**Station t**	
	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT
IOA	0.71	0.75	0.70	0.72	0.71	0.75	0.72	0.73	0.70	0.75
R	0.51	0.60	0.51	0.56	0.53	0.59	0.53	0.56	0.55	0.59
RMSE	16.85	12.91	16.73	12.71	16.47	12.65	17.07	14.60	15.80	11.07
MB	1.74	1.54	1.52	−0.88	1.48	−0.71	0.65	−1.15	3.48	0.80
PBIAS	7.67	6.81	6.66	−3.85	6.45	−3.09	2.71	−4.79	16.57	3.82
Metrics	**Station u**		**Station v**		**Station w**		**Station x**		**Station y**	
	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT	3-D CTM	HAT
IOA	0.74	0.75	0.71	0.74	0.71	0.75	0.71	0.76	0.73	0.76
R	0.56	0.61	0.51	0.61	0.51	0.61	0.52	0.61	0.54	0.62
RMSE	17.10	14.73	17.45	13.64	15.86	11.79	16.83	13.14	16.56	13.36
MB	2.85	2.83	2.03	1.01	1.57	1.91	0.86	0.96	−2.83	−0.40
PBIAS	13.19	13.08	8.89	4.42	7.57	9.18	3.70	4.13	−11.13	−1.56

* IOA and R are dimensionless; the units for RMSE and MB are $\mu g/m^3$; the unit for PBIAS is %.

Further analysis of the error and bias metrics reinforces the superior performance of the HAT model. The RMSE values for the HAT predictions ranged from 11.07 $\mu g/m^3$ to 14.73 $\mu g/m^3$, with a mean RMSE of 12.82 $\mu g/m^3$, while the 3-D CTM predictions exhibited higher RMSE values ranging from 15.56 $\mu g/m^3$ to 17.45 $\mu g/m^3$, with a mean RMSE of 16.45 $\mu g/m^3$. This significant difference in RMSE highlights the HAT's superior performance. Additionally, the HAT model demonstrated significantly lower bias, with a mean MB of 0.26 $\mu g/m^3$ compared to 1.47 $\mu g/m^3$ for the 3-D CTM, representing an 82.59% reduction in bias. This reduction underscores the HAT model's ability to track true $PM_{2.5}$ levels closely, enhancing its practical reliability. For PBIAS, the model demonstrated values ranging from −5.64% to 13.08%, with an average of 1.13%, whereas the 3-D CTM showed greater variability, ranging from −11.13% to 18.73%, with an average of 6.50%. These results confirm the HAT model's robustness and reliability in minimizing both errors and systematic biases for $PM_{2.5}$ prediction.

Figure 5 provides a box plot comparison of $PM_{2.5}$ prediction performance metrics, including IOA, R, RMSE, MB, and PBIAS, for the 3-D CTM and HAT across the 25 prediction points. The box plots offer a visual representation of the central tendency, variability, and distribution for each metric. The results demonstrate that the HAT model achieves higher IOA and R values, as indicated by the upward shift in medians and reduced variability compared to the 3-D CTM. For RMSE, the box plots show significantly lower error values for the HAT model, suggesting its greater predictive precision. Similarly, the MB and PBIAS box plots reveal smaller biases for the HAT model, with median values closer to zero and reduced variability, highlighting its superior stability and accuracy.

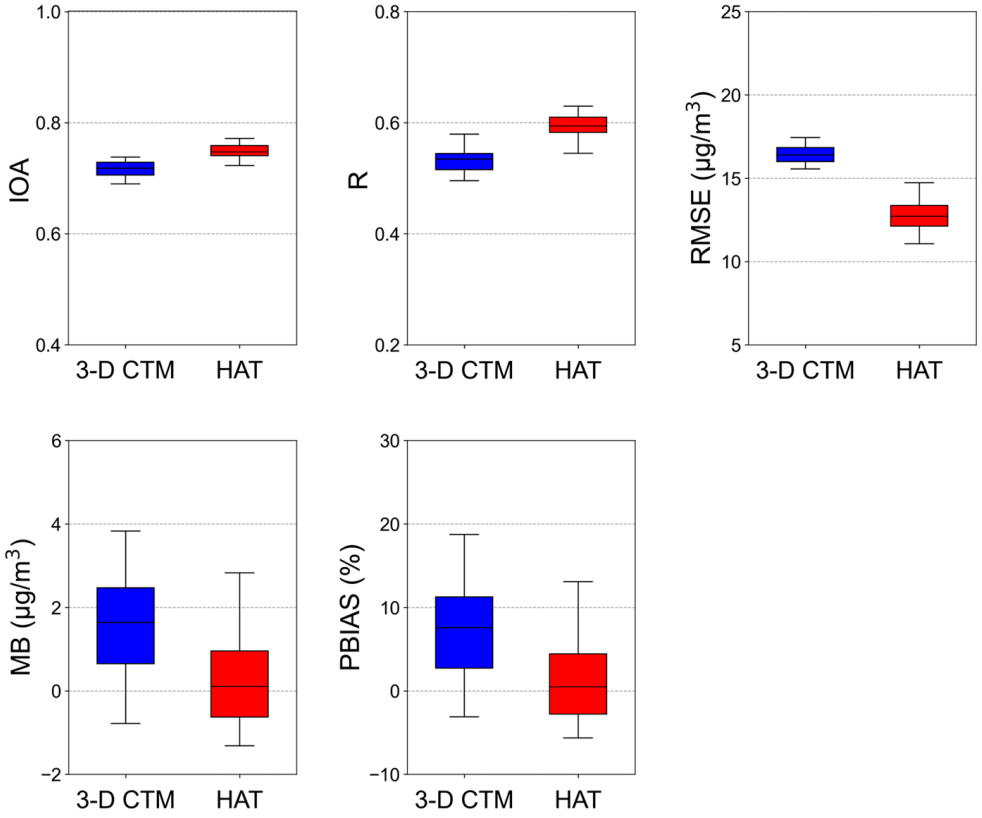

Figure 5. Box plot comparison of $PM_{2.5}$ prediction performance metrics across prediction points.

Compared to Kim et al. [16], who employed an LSTM-based approach for $PM_{2.5}$ predictions in Seoul and reported MB values ranging from -1.09 μg/m^3 to -1.33 μg/m^3, the HAT model demonstrated superior performance across multiple evaluation metrics, achieving a mean MB of 0.26 μg/m^3, indicating greater accuracy and lower bias. While both studies utilized the same input features, Kim et al. [16] developed individual LSTM models for each prediction point, limiting their approach to specific locations. In contrast, the HAT model, designed as a single unified framework, was assessed across 25 prediction points spanning diverse geographical areas, highlighting its robustness and scalability. Despite the LSTM model's optimization for individual sites, it exhibited relatively large negative biases when applied across multiple prediction points, whereas the HAT model consistently minimized systematic errors across all locations. These findings establish the HAT model as a versatile and scalable solution for daily $PM_{2.5}$ predictions, effectively generalizing across multiple locations while maintaining high accuracy.

3.2. Characteristics of HAT Prediction

To investigate the detailed characteristics of the HAT model, we analyzed the diurnal variation in $PM_{2.5}$ predictions, with the results summarized in Figure 6. Significant

differences were found in the diurnal patterns of $PM_{2.5}$ predictions between the HAT and 3-D CTM. Consistent with the previous comparative analysis, the HAT predictions consistently aligned with the observed $PM_{2.5}$ throughout the day. In contrast, 3-D CTM-based predictions exhibited significant diurnal variability, following a sinusoidal pattern, with an increase until approximately 08:00 local standard time (LST; UTC + 09:00), a decrease until 17:00 LST, and a subsequent rise thereafter from 17:00 LST onward. This pattern reflects the complex interactions among meteorological conditions, emissions, and atmospheric physicochemical processes. The morning increase in $PM_{2.5}$ is primarily driven by elevated emissions from anthropogenic activities. As daytime solar radiation intensifies, rising surface temperature increases wind speed and mixing height, promoting turbulent mixing and contributing to a gradual $PM_{2.5}$ decrease. In contrast, the evening rise is linked to a decrease in mixing height, which limits vertical dispersion and results in the accumulation of $PM_{2.5}$ near the surface.

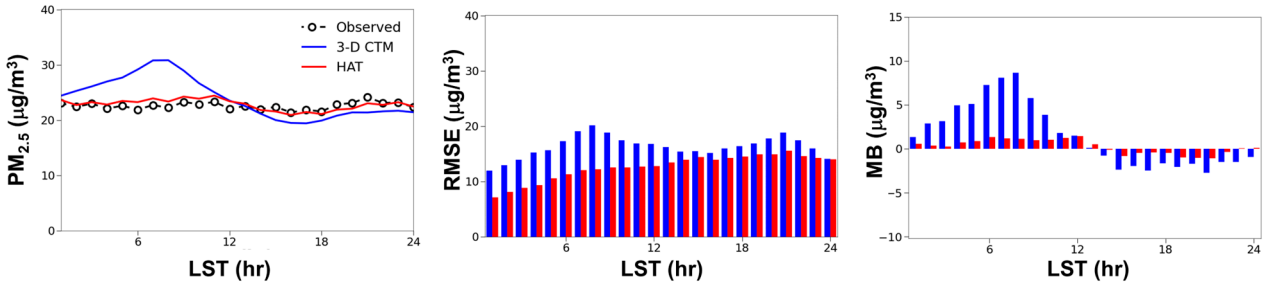

Figure 6. Diurnal variation of $PM_{2.5}$ predictions and their errors and biases.

The RMSE patterns in the HAT and 3-D CTM predictions revealed distinct diurnal trends. For the HAT, the RMSE increased as the time gap between the input features and the corresponding true values lengthened. This behavior is typical in DNN-based time-series predictions, where a longer time interval between the input sequence and the prediction timepoint diminishes the correlation between the input data and the true values, leading to a higher RMSE. In contrast, the 3-D CTM predictions exhibited both increases and decreases in RMSE throughout the day, indicating greater variability in the model's performance over time. These fluctuations were attributed to daily variations in meteorological conditions, emissions, and physicochemical processes in the 3-D CTM predictions.

Both models exhibited similar diurnal patterns in the MB, although there were differences in the magnitude of the biases. Positive biases were observed until 13:00 LST, after which both predictions showed negative biases. At 08:00 LST, the time of maximum positive bias for the 3-D CTM, the MB of the HAT was 87.06% smaller than that of the 3-D CTM. Similarly, at 21:00 LST, when the negative bias reached its peak, the MB of HAT was 60.34% smaller than that of 3-D CTM. These results indicate that, while both models exhibited similar diurnal trends in bias, the HAT consistently provided a more accurate representation of $PM_{2.5}$ levels throughout the day. Although the general patterns of over- and underestimation were comparable, the HAT's smaller bias magnitudes emphasize its superior precision and closer alignment with observed $PM_{2.5}$, underscoring its enhanced prediction capability.

3.3. Feature Importance

Understanding the interaction between input features and predicted outcomes in complex DNNs is challenging due to the nonlinear relationships between the input data and the network components. DNNs process inputs through multiple layers, where each layer sequentially applies nonlinear transformations to the input vectors, generating intermediate features at each stage. The nonlinearity introduced by the intricate structure

of DNN operations complicates the assessment of how individual features influence the final predictions. Furthermore, these interactions make it difficult to evaluate the true importance of any single feature, as the effect of one feature on the prediction is often intertwined with the effects of others.

To address this, we employed the permutation feature importance method, which involves randomly shuffling individual input features and observing the resulting change in model performance. This approach allows for isolating the individual contribution of each feature by breaking down the complex interactions among them, effectively enabling the estimation of each feature's independent importance. A notable decrease in performance after perturbation indicates the feature's significance, while a minimal decrease suggests a lower contribution. Conversely, an improved performance implies that the feature may be unnecessary. IOA was used as the baseline to evaluate feature importance. The permutation importance was estimated as follows:

$$FI_i = \frac{\Delta IOA_{i,\ P}}{IOA_{w/o,\ P}} \times 100 \tag{18}$$

here FI_i refers to the importance of feature i, $IOA_{w/o,P}$ represents the accuracy of the HAT predictions without feature permutation and $\Delta IOA_{i,P}$ denotes the change in the accuracy after the permutation of feature i, respectively. By reflecting the percentage drop in accuracy, this metric highlights how critical each feature is to the model's overall predictive capability. The findings from the importance estimation are presented in Figure 7.

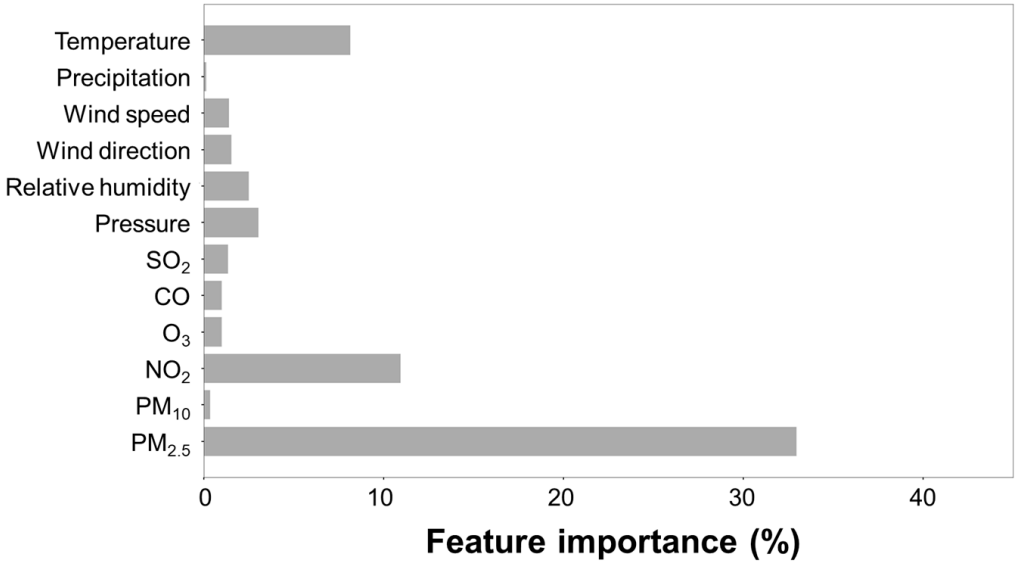

Figure 7. Summary of the feature importance assessment.

As illustrated in Figure 7, the permutation importance of all input features was found to be positive, indicating that all variables are essential for daily $PM_{2.5}$ predictions. Among the features, the previous day's $PM_{2.5}$ and NO_2 exhibited a significant influence on the next day's $PM_{2.5}$. Specifically, the $PM_{2.5}$ from the previous day were identified as the most influential feature, with an importance value of 32.96%. This considerable influence can be attributed to the temporal persistence of $PM_{2.5}$, where concentrations from the previous day show a strong correlation with those of the following day.

Among the meteorological features, temperature and pressure were found to be the most influential, with importance values of 8.13% and 2.03%, respectively. In South Korea, $PM_{2.5}$ exhibits significant seasonal variations, typically peaking in winter and diminishing in summer. This seasonal pattern is primarily driven by two factors: (i) increased

anthropogenic emissions, particularly due to higher fossil fuel consumption for domestic heating, and (ii) enhanced long-range transport of air pollutants from neighboring countries, facilitated by intensified westerly winds. Temperature and pressure, as key indicators of seasonality, contributed substantially to the model's accuracy by reflecting seasonal variations that influence $PM_{2.5}$. Consequently, the importance of these variables was relatively higher than that of other meteorological features, highlighting their critical role in $PM_{2.5}$ predictions.

3.4. Advantages of HAT

To thoroughly investigate the advantages of the proposed HAT, we conducted a comparative analysis by evaluating its $PM_{2.5}$ predictions against those produced by an LSTM-based approach. This analysis assessed their respective capabilities to handle complex temporal patterns and atmospheric environmental variability, both of which are critical for accurate $PM_{2.5}$ predictions. The LSTM, renowned for its ability to capture sequential dependencies in time-series data, served as a benchmark to assess the capability and robustness of the HAT relative to the traditional DNN approaches. The LSTM architecture consisted of three sequential LSTM layers with 1024, 512, and 256 hidden units, designed to capture temporal dependencies progressively. A fully connected output layer with 24 nodes was employed to map the extracted latent features to observed $PM_{2.5}$ via a nonlinear transformation. This multilayer configuration was specifically designed to capture both short- and long-term signals, reflecting the inherent temporal scales of $PM_{2.5}$ dynamics.

As illustrated in Figure 8, the HAT outperformed the LSTM, exhibiting an 18.50% smaller error and 95.91% lower bias. These results are consistent with previous research demonstrating that transformer-based models, such as the HAT, tend to achieve more accurate $PM_{2.5}$ predictions compared to the LSTM-based models [31–35]. While the LSTM demonstrated similar prediction accuracy to the 3-D CTM, it exhibited a relatively large positive bias of 6.35 $\mu g/m^3$. This bias can be attributed to gradient vanishing during the LSTM's 5-year training period, as well as abnormal atmospheric conditions during the prediction period, particularly those influenced by El Niño. Gradient vanishing is a well-known issue in LSTM models, where gradients become excessively small during backpropagation, significantly impeding the model's ability to learn long-term dependencies. This issue is particularly likely to arise when training on long sequences, as in the case of the 5-year dataset employed in this study. The extended temporal span exacerbates the difficulty of effectively updating weights for earlier time steps, which in turn limits the LSTM's capacity to capture intricate temporal patterns over such prolonged training periods. Additionally, the period was influenced by the El Niño phenomenon, which caused significant atmospheric anomalies. During this time, the Oceanic Niño Index (ONI) ranged from 0.7 °C to 1.8 °C, indicating strong El Niño activity. El Niño events are known to significantly impact $PM_{2.5}$ variability in South Korea, particularly during the winter months [42,43]. These events lead to warmer winter temperatures in East Asia, weaken the East Asian winter monsoon, and consequently influence $PM_{2.5}$ concentrations. Despite these challenges, the HAT model exhibited substantially superior predictive performance compared to the LSTM. Its transformer-based architecture appears to handle environmental anomalies and extended historical data more effectively, leveraging self-attention mechanisms to maintain relevant information across longer time frames. This highlights the robustness of the HAT in accurately predicting $PM_{2.5}$, even under unfamiliar and abnormal atmospheric conditions associated with El Niño.

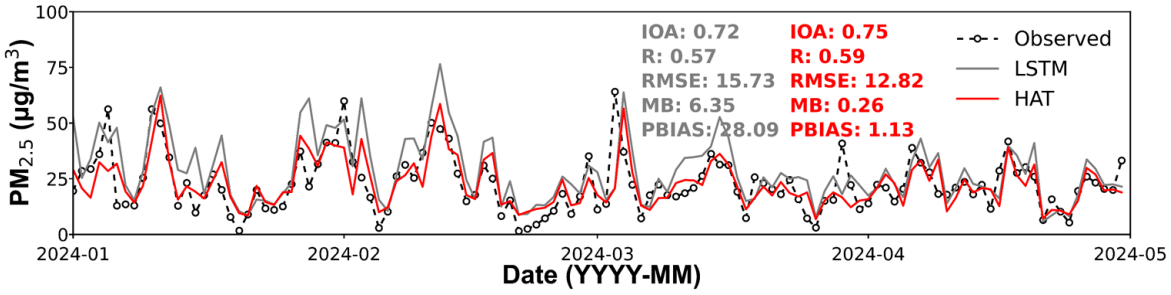

Figure 8. Comparative performance analysis of the HAT and LSTM models. Black open circles with dashed lines represent observed $PM_{2.5}$, while grey and red lines represent the LSTM and HAT $PM_{2.5}$ predictions, respectively.

4. Conclusions

In this study, we developed a HAT model to accurately predict daily $PM_{2.5}$ in Seoul. Compared to traditional 3-D CTM-based methods, the HAT offers a more cost-efficient and accurate approach to air quality prediction in a data-driven manner. While 3-D CTMs require labor-intensive preprocessing—such as preparing emissions, obtaining boundary conditions, and running numerical weather prediction models—the HAT model bypasses these complexities, providing a more streamlined and efficient predictive framework. Additionally, the 3-D CTM relies on resource-intensive parallel computing systems for operational air quality forecasting, involving large-scale operations and complex computations to account for sophisticated atmospheric processes. In contrast, the HAT model significantly reduces computational costs by efficiently processing observational datasets and generating next-day predictions within seconds without requiring such detailed atmospheric considerations. Given these advantages, the HAT serves as a practical alternative for researchers, policymakers, and institutions facing financial and technical challenges in implementing 3-D CTM-based forecasting systems. Its simplicity, data-driven nature, and accessibility as a public resource make it an ideal tool for enhancing air quality forecasting capabilities and enabling widespread adoption without reliance on costly computational infrastructure.

The predictive performance of the HAT was evaluated by comparing its predictions with both ground observations and those from the 3-D CTM. The results indicated that the HAT model achieved a 4.60% higher IOA compared to the 3-D CTM, demonstrating a superior ability to capture the temporal dynamics of $PM_{2.5}$. Specifically, the HAT model's RMSE and MB were 22.09% and 82.59% lower, respectively, than those of the 3-D CTM. Furthermore, the HAT effectively captured the diurnal variation of $PM_{2.5}$, aligning more closely with ground observations throughout the day compared to the 3-D CTM. Additionally, comparisons with the LSTM model highlighted the robustness of the HAT, maintaining strong performance even under challenging atmospheric conditions, such as during El Niño events. These findings suggest that the HAT can generalize effectively across varying timeframes and environmental conditions, making it a reliable tool for $PM_{2.5}$ prediction.

However, the current HAT-based approach exhibits relatively large errors in predicting high $PM_{2.5}$. These errors are primarily attributed to the insufficient representation of training samples that correspond to high $PM_{2.5}$ events, which hampers the model's ability to generalize under such conditions. Additionally, abnormal atmospheric patterns, such as those induced by El Niño, introduce additional variability and complexity, compounding these challenges. To address these limitations, future research will focus on securing additional data specific to high $PM_{2.5}$ events, thereby improving the representation of extreme conditions in the training dataset. Additionally, leveraging multiyear datasets will enhance the model's robustness and adaptability to diverse atmospheric conditions. These efforts aim to optimize the model's performance and mitigate its current shortcomings.

Furthermore, the framework developed in this study will be expanded to include daily predictions of various air pollutants, thereby broadening its applicability and providing a more comprehensive tool for air quality prediction. Despite its current limitations, the HAT model demonstrates significant potential as a practical alternative to resource-intensive 3-D CTM-based methods for operational air quality forecasting.

Author Contributions: Conceptualization, H.S.K.; methodology, H.S.K. and K.M.H.; software, H.S.K. and J.Y.; validation H.S.K., N.Y. and T.C.; writing—original draft preparation H.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2022R1I1A1A01066083).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: We obtained KMA ASOS and NIER AIR KOREA datasets from respective official data archives. This work appreciates the institutions that provided useful observational data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Brook, R.D.; Rajagopalan, S.; Pope, C.A.; Brook, J.R.; Bhatnagar, A.; Diez-Roux, A.V.; Holguin, F.; Hong, Y.; Luepker, R.V.; Mittleman, M.A.; et al. Particulate matter air pollution and cardiovascular disease: An Update to the Scientific Statement from the American Heart Association. *Circulation* **2010**, *121*, 2331–2378. [CrossRef]
2. Crouse, D.L.; Peters, P.A.; Hystad, P.; Brook, J.R.; van Donkelaar, A.; Martin, R.V.; Villeneuve, P.J.; Jerrett, M.; Goldberg, M.S.; Arden Pope, C.; et al. Ambient $PM_{2.5}$, O_3, and NO_2 exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environ. Health Perspect.* **2015**, *123*, 1180–1186. [CrossRef]
3. Xing, Y.-F.; Xu, Y.-H.; Shi, M.-H.; Lian, Y.-X. The impact of $PM_{2.5}$ on the human respiratory system. *J. Thorac. Dis.* **2016**, *8*, E69–E74.
4. Dorkery, D.W.; Schwartz, J.; Spengler, J.D. Air pollution and daily mortality: Associations with particles and acid aerosols. *Environ. Res.* **1992**, *59*, 362–373. [CrossRef]
5. Pope, C.A., III; Dorkey, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manage. Assoc.* **2006**, *56*, 709–742. [CrossRef]
6. Leiva, G.M.A.; Santibañez, D.A.; Ibarra, E.S.; Matus, C.P.; Seguel, R. A five-year study of particulate matter ($PM_{2.5}$) and cerebrovascular diseases. *Environ. Pollut.* **2013**, *181*, 1–6. [CrossRef] [PubMed]
7. Pun, V.C.; Kazemiparkouhi, F.; Manjourides, J.; Suh, H.H. Long-term $PM_{2.5}$ exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *Am. J. Epidemiol.* **2017**, *186*, 961–969. [CrossRef]
8. Berge, E.; Huang, H.-C.; Chang, J.; Liu, T.-H. A study of the importance of initial conditions for photochemical oxidant modeling. *J. Geophys. Res.-Atmos.* **2001**, *106*, 1347–1363. [CrossRef]
9. Liu, T.-H.; Jeng, F.-T.; Huang, H.-C.; Berger, E.; Chang, J.S. Influences of initial conditions and boundary conditions on regional and urban scale Eulerian air quality transport model simulations. *Chemosphere-Glob. Change Sci.* **2001**, *3*, 175–183. [CrossRef]
10. Holloway, T.; Spak, S.N.; Barker, D.; Bretl, M.; Moberg, C.; Hayhoe, K.; Van Dorn, J.; Wuebbles, D. Change in ozone air pollution over Chicago associated with global climate change. *J. Geophys. Res.-Atmos.* **2008**, *113*, D22306. [CrossRef]
11. Han, K.M.; Lee, C.K.; Lee, J.; Kim, J.; Song, C.H. A comparison study between model-predicted and OMI-retrieved tropospheric NO_2 columns over the Korean peninsula. *Atmos. Environ.* **2011**, *45*, 2962–2971. [CrossRef]
12. McKendry, I.G. Evaluation of artificial neural networks for fine particulate pollution (PM_{10} and $PM_{2.5}$) forecasting. *J. Air Waste Manag. Assoc.* **2002**, *52*, 1096–1101. [CrossRef]
13. Lu, W.Z.; Fan, H.Y.; Lo, S.M. Application of evolutionary neural network method in predicting pollutant levels in downtown area of Hong Kong. *Neurocomputing* **2003**, *51*, 387–400. [CrossRef]
14. Pozza, S.A.; Lima, E.P. Time series analysis of $PM_{2.5}$ and $PM_{10-2.5}$ mass concentration in the city of Sao Carlos, Brazil. *Int. J. Environ. Pollut.* **2010**, *41*, 90–108. [CrossRef]

15. Tsai, Y.; Zeng, Y.; Chang, Y. Air pollution forecasting using RNN with LSTM. In Proceedings of the 2018 IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Athens, Greece, 12–15 August 2018.
16. Kim, H.S.; Park, I.; Song, C.H.; Lee, K.; Yun, J.W.; Kim, H.K.; Jeon, M.; Lee, J.; Han, K.M. Development of a daily PM_{10} and $PM_{2.5}$ prediction system using a deep long short-term memory neural network model. *Atmos. Chem. Phys.* **2019**, *19*, 12935–12951. [CrossRef]
17. Chae, S.; Shin, J.; Kwon, S.; Lee, S.; Kang, S.; Lee, D. PM_{10} and $PM_{2.5}$ real-time prediction models using an interpolated convolutional neural network. *Sci. Rep.* **2021**, *11*, 11952. [CrossRef] [PubMed]
18. Chang-Hoi, H.; Park, I.; Oh, H.-R.; Gim, H.-J.; Hur, S.-K.; Kim, J.; Choi, D.-R. Development of a $PM_{2.5}$ prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea. *Atmos. Environ.* **2021**, *245*, 118021. [CrossRef]
19. Li, T.; Hua, M.; Wu, X. A hybrid CNN-LSTM model for forecasting particulate matter ($PM_{2.5}$). *IEEE Access* **2020**, *8*, 26933–26940. [CrossRef]
20. Kim, H.S.; Han, K.M.; Yu, J.; Kim, J.; Kim, K.; Kim, H. Development of a CNN+LSTM hybrid neural network for daily $PM_{2.5}$ Prediction. *Atmosphere* **2022**, *13*, 2124. [CrossRef]
21. Hopfield, J.J. Hopfield network. *Scholarpedia* **2007**, *2*, 1977. [CrossRef]
22. Hu, Y.; Huber, A.; Anumula, J.; Liu, S.-C. Overcoming the vanishing gradient problem in plain recurrent networks. *arXiv* **2018**, arXiv:1801.06105.
23. Singh, D.; Choi, Y.; Park, J.; Salman, A.K.; Sayeed, A.; Song, C.H. Deep-BCSI: A deep learning-based framework for bias correction and spatial imputation of $PM_{2.5}$ concentrations in South Korea. *Atmos. Res.* **2024**, *301*, 107283. [CrossRef]
24. Ding, C.; Wang, G.Z.; Zhang, X.Y.; Liu, Q.; Liu, X.D. A hybrid CNN-LSTM model for predicting $PM_{2.5}$ in Beijing based on spatiotemporal correlation. *Environ. Ecol. Stat.* **2021**, *28*, 503–522. [CrossRef]
25. Li, S.; Xie, G.; Ren, J.; Guo, L.; Yang, Y.; Xu, X. Urban $PM_{2.5}$ concentration prediction via attention-based CNN–LSTM. *Appl. Sci.* **2020**, *10*, 1953. [CrossRef]
26. Kolen, J.F.; Kremer, S.C. Gradient flow in recurrent nets: The difficulty of learning longterm dependencies. In *A Field Guide to Dynamical Recurrent Networks*; Wiley-IEEE Press: Hoboken, NJ, USA, 2010; pp. 237–243.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
29. Tong, W.T.; Limperis, J.; Hamza-Lup, F.; Xu, Y.; Li, L.X. Robust transformer-based model for spatiotemporal $PM_{2.5}$ prediction in California. *Earth Sci. Inform.* **2023**, *17*, 315–328. [CrossRef]
30. Gao, Z.; Mo, X.; Li, H. Prediction of $PM_{2.5}$ concentration based on deep learning, multi-objective optimization, and ensemble forecast. *Sustainability* **2024**, *16*, 4643. [CrossRef]
31. Cui, B.; Liu, M.; Li, S.; Jin, Z.; Zeng, Y.; Lin, X. Deep learning methods for atmospheric $PM_{2.5}$ prediction: A comparative study of transformer and CNN-LSTM-attention. *Atmos. Pollut. Res.* **2023**, *14*, 101833. [CrossRef]
32. Wang, H.; Zhang, L.; Wu, R. MSAFormer: A transformer-based model for $PM_{2.5}$ prediction leveraging sparse autoencoding of multi-site meteorological features in Urban Areas. *Atmosphere* **2023**, *14*, 1294. [CrossRef]
33. Rai, V.; Kumar, S.; Sihgh, T.; Kapoor, R.P. $PM_{2.5}$ level forecasting using transformer-based model. In Proceedings of the 3rd International Conference on Advance Computing and Innovative Technologies in Engineering, Greater Noida, India, 12–13 May 2023.
34. Dai, Z.; Ren, G.; Jin, Y.; Zhang, J. Research on $PM_{2.5}$ concentration prediction based on transformer. In Proceedings of the 7th International Symposium on Big Data and Applied Statistics, Beijing, China, 8–10 March 2024.
35. Zou, R.; Huang, H.; Lu, X.; Zeng, F.; Ren, C.; Wang, W.; Zhou, L.; Dai, X. PD-LL-Transformer: An hourly $PM_{2.5}$ forecasting method over the Yangtze River Delta Urban agglomeration, China. *Remote Sens.* **2024**, *16*, 1915. [CrossRef]
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
37. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 3–8 May 2015.
38. Guenther, A.; Karl, T.; Harley, P.; Wiedinmyer, C.; Palmer, P.I.; Geron, C. Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmos. Chem. Phys.* **2006**, *6*, 3181–3210. [CrossRef]
39. Salman, A.K.; Choi, Y.; Singh, D.; Kayastha, S.G.; Dimri, R.; Park, J. Temporal CNN-based 72-h ozone forecasting in South Korea: Explainability and uncertainty quantification. *Atmos. Environ.* **2025**, *343*, 120987. [CrossRef]

40. Koo, J.-S.; Wang, K.-H.; Yun, H.-Y.; Kwon, H.-Y.; Koo, Y.-S. Development of $PM_{2.5}$ Forecast Model Combining ConvLSTM and DNN in Seoul. *Atmosphere* **2024**, *15*, 1276. [CrossRef]
41. Tao, H.; Abba, S.I.; Al-Areeq, A.M.; Tangang, F.; Samantaray, S.; Sahoo, A.; Siqueira, H.V.; Maroufpoor, S.; Demir, V.; Bokde, N.D.; et al. Hybridized artificial intelligence models with nature-inspired algorithms for river flow modeling: A comprehensive review, assessment, and possible future research directions. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107559. [CrossRef]
42. Jeong, J.I.; Park, R.J.; Yeh, S.-W. Dissimilar effects of two El Niño types on $PM_{2.5}$ concentrations in East Asia. *Environ. Pollut.* **2018**, *242*, 1395–1403. [CrossRef]
43. Jeong, J.I.; Park, R.J.; Song, C.-K.; Yeh, S.-W.; Woo, J.-H. Quantitative analysis of winter $PM_{2.5}$ reduction in South Korea, 2019/20 to 2021/22: Contributions of meteorology and emissions. *Sci. Total Environ.* **2024**, *907*, 168179. [CrossRef]