atmosphere

MDPI

Article

Air Pollutant Concentration Forecasting with WTMP: Wavelet Transform-Based Multilayer Perceptron

Xiaoling Wang [1], Liangzhao Tao [1,*], Mingliang Fu [2,*] and Qi Wang [3]

[1] School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou 310018, China
[2] State Environmental Protection Key Laboratory of Vehicle Emission Control and Simulation, Chinese Research Academy of Environmental Sciences, Beijing 100012, China
[3] College of Life and Environmental Sciences, Wenzhou University, Wenzhou 325000, China
* Correspondence: 22020090021@pop.zjgsu.edu.cn (L.T.); fumingliang160@163.com (M.F.)

Abstract: Atmospheric pollutants' real-time changes and the internal interactions among various data make it challenging to efficiently predict concentration variations. In order to extract more information from the time series of pollutants and improve the accuracy of prediction models, we propose a type of Multilayer Perceptron model based on wavelet decomposition, named Wavelet Transform-based Multilayer Perceptron (WTMP) model. This model decomposes pollutant data through overlapping discrete wavelet transforms to extract non-stationarity and nonlinear dependencies in the time series. It combines the decomposed data with static covariate information such as data collection time and inputs them into an improved Multilayer Perceptron (MLP) model, reconstructing and outputting the prediction results. Finally, the model is validated using atmospheric pollutant data collected at a specific location in Ruian City, Zhejiang Province, China. The results indicate that the model performs well with minimal prediction errors.

Keywords: atmospheric pollutants; wavelet decomposition; multi-layer perceptron model

check for updates

Citation: Wang, X.; Tao, L.; Fu, M.; Wang, Q. Air Pollutant Concentration Forecasting with WTMP: Wavelet Transform-Based Multilayer Perceptron. *Atmosphere* **2024**, *15*, 1296. https://doi.org/10.3390/atmos15111296

Academic Editors: Antonio Donateo, Zhenyi Xu, Changfa Tao

Received: 3 September 2024
Revised: 9 October 2024
Accepted: 14 October 2024
Published: 28 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, the escalating issue of environmental pollution has brought the subject of air pollution to the forefront of global concern [1], especially at the United Nations Climate Change Conferences. The primary atmospheric pollutants include six categories: nitrogen oxides, sulfur oxides, carbon oxides, ozone, PM10, and PM2.5 Specifically, PM2.5 refers to particulate matter with a diameter less than 2.5 microns, which are so minuscule and lightweight that they can linger in the atmosphere for extended periods. When inhaled, these particles can bypass the nasal cavity, reaching the throat and even the lungs [2], potentially leading to conditions such as bronchitis, asthma, cardiovascular diseases, and lung cancer, thereby posing a significant risk to human health. In 2018, research by Heft-Neal et al. [3] suggested that PM2.5 levels above the minimum exposure threshold are a primary cause of 22% of infant mortality across 30 countries. Additionally, various atmospheric pollutants can irritate the respiratory tract, predisposing individuals to respiratory illnesses. Prolonged exposure to polluted environments may result in impaired vascular endothelial function and an array of cardiovascular diseases. Fine particulate matter in the atmosphere can also impact the nervous system, elevating the risk of neurological disorders such as Alzheimer's and Parkinson's diseases.

Given the profound impact of air pollution on living environments and human health, research into air pollution forecasting has become increasingly vital [4]. It is recognized as a crucial aspect of environmental protection and a key instrument in guiding scientific decision-making for air pollution alerts and control measures. Numerous cities have established air quality monitoring stations to provide real-time surveillance of PM2.5 and other atmospheric pollutants.

The composition and concentration of various gas pollutants form a time series, characterized by chronological order. By analyzing historical data, it is possible to identify dynamic patterns and predict future values. This capability is crucial for forecasting problems involving time-dependent components. This paper focuses on analyzing and extracting patterns from the concentration data of air pollutants within a specific time period to predict future air quality levels. Researchers have discovered that, in addition to other atmospheric pollutants, climatic factors such as wind speed, wind direction, and relative humidity significantly impact the prediction of future concentrations [5]. Therefore, it is essential to extract these complex features, determine their correlations, and construct a time series model to foresee potential future changes.

With the continuous advancement of science and technology and the widespread use of various sensors, air quality monitoring networks have been established nationwide. These networks, equipped with sensors covering multiple monitoring points, facilitate the collection of time series data, making it increasingly easier to obtain various data. Consequently, there is a higher demand for accurate results from time series predictions. Multivariate analysis and precise forecasting of time series data are becoming increasingly important in time prediction. Currently, time series forecasting techniques are a significant research focus in many fields, including weather prediction, disease prevention, pollutant detection, and stock market forecasting [6]. They are also relevant in environmental management and forecasting, regional development planning, and other areas.

In the evolution of air quality prediction, numerical modeling methods have seen significant advancements. Early numerical models primarily relied on simple physical and chemical processes to forecast the dispersion of pollutants. For instance, in 2004, Arystanbekova K et al. [7] proposed the Gaussian plume model based on the Gaussian distribution characteristics of atmospheric pollutant concentrations. While these models offered an initial understanding of pollutant behavior, they often had limited capabilities when dealing with complex terrain and meteorological conditions.

With the increase in computational power, modern numerical models such as climate models and regional air quality models (like the Community Multiscale Air Quality (CMAQ) model and the Weather Research and Forecasting (WRF)-Chem model) have incorporated more complex chemical reactions and meteorological factors. These models can more accurately simulate the generation, transformation, and transport of pollutants, providing high-resolution forecasts. However, they typically require a substantial amount of input data and come with high computational costs, and in practical applications, they may face challenges in data acquisition and model parameterization.

In contrast, data-driven models, such as machine learning methods, train on historical observational data and can quickly generate predictions, often performing well when there is an abundance of data. However, they lack an in-depth understanding of physical processes, which may lead to suboptimal performance in the face of extreme weather or new pollutants.

With the continuous development of computer technology, traditional time series forecasting methods, often used for short-term univariate predictions, struggle with nonlinear relationships and multivariate, complex data. Consequently, machine learning and data mining techniques have been introduced to address these challenges [8]. These methods analyze data patterns, extract features, and capture correlations through training to predict target events. Machine learning-based models excel at fitting nonlinear time series data but often overlook historical temporal dependencies, resulting in high computational complexity and potential local optimization issues. In contrast, deep learning methods are highly effective in large-scale data feature extraction and model construction [9]. Deep neural networks, with their multiple hidden layers, can identify complex nonlinear patterns and require large datasets for accurate predictions.Recurrent Neural Network (RNN) models address the temporal characteristics of time series data [10], offering short-term memory capabilities. They are widely used in fields such as financial forecasting and power load prediction. Long Short-Term Memory (LSTM) networks, a variant of RNNs, were pro-

posed by Hochreiter et al. [11] to address gradient vanishing in long-term dependencies. For example, Li and colleagues developed an air quality prediction model using LSTM networks [12], which outperformed ARIMA [13] and SVR [14] models. Following this, Wu and colleagues [15] integrated LSTM with quadratic decomposition and optimization techniques to create a hybrid forecasting model. Additionally, Tao and colleagues proposed a model combining 1D-CNNs and Bi-GRU [9] for predicting the Air Pollution Index. Building on the success of CNNs in image analysis, Dong and colleagues hypothesized its potential for deciphering the spatial patterns of air pollutants. Wu and colleagues enhanced prediction accuracy by applying CNN to multi-site air quality data [16].

To overcome the limitations of single networks, various combinations of deep learning networks have been explored. By integrating CNN's feature extraction capabilities with LSTM's proficiency in handling time series data, Kuo and colleagues [17] introduced the CNN-LSTM model for PM2.5 forecasting. Similarly, Pak and colleagues developed a CNN-LSTM model for ozone prediction [18], significantly improving predictive capabilities. In recent decades, the complexity of factors influencing air pollution predictions, combined with the widespread use of various sensors, has led to the establishment of air quality monitoring networks across the country. This has enabled the collection of comprehensive time series data, providing valuable insights for developing more effective and accurate industrial pollutant index prediction models. Qi and colleagues combined LSTM networks with Graph Convolutional Networks (GCN) to predict PM2.5 concentrations in the air [19]. They proposed a hybrid GC-LSTM model and validated its effectiveness and feasibility by evaluating metrics such as recall rate, false alarm rate, and correlation coefficient matrices.

In the domain of sequence modeling such as language and vision, Transformers [20] have emerged as the most successful deep learning architecture, which utilize multi-head attention and positional encoding to significantly enhance the model's ability to capture complex relationships within data [21]. By leveraging multiple attention heads, the Transformer can focus on different parts of the input sequence simultaneously, leading to a richer representation of the data [22]. The inclusion of positional encoding allows the model to understand the order of the sequence, which is crucial for tasks involving time series data where temporal relationships are important.

An advanced Transformer-based model, known as Informer [23], is specifically designed for long-sequence prediction. It introduces three key innovations to improve efficiency and reduce memory consumption in long-sequence forecasting: ProbSparse self-attention, which selects the most relevant attention scores to focus on, thereby reducing computational complexity; self-attention distillation, which distills the attention maps to retain the most important information while discarding redundancy; and generative decoding, which enhances the model's ability to generate long sequences by iteratively predicting future values.

The model, Autoformer [24], proposes a depth-wise decomposition architecture and a self-correlation mechanism to more effectively forecast time series data. The depth-wise decomposition architecture separates trend and seasonal components of the time series, allowing the model to capture both long-term trends and short-term fluctuations more accurately. The self-correlation mechanism leverages the inherent autocorrelation in time series data to improve prediction accuracy by identifying and utilizing repeating patterns within the data.

The model incorporates a frequency-enhanced structure, known as FEDFormer [25], to enhance the accuracy of time series forecasting by better capturing frequency components. This structure allows the model to better handle periodic variations in time series data, excelling in capturing both long-term and short-term patterns. Pyraformer [26] uses a pyramidal self-attention mechanism, which has linear complexity and can attend to different granularities of time information. This multi-granularity attention allows Pyraformer to flexibly handle complex time series data while maintaining computational efficiency.

Recently, the DLinear model [27] has been introduced as a straightforward linear model that, contrary to expectations, surpasses the performance of numerous Transformer-

based models. DLinear adeptly acquires a direct linear relationship between the context and the forecast horizon, highlighting the limitations of sub-quadratic approximations used in self-attention mechanisms. This model focuses on the simplicity and interpretability of linear relationships, demonstrating that in many cases, complex attention mechanisms may not be necessary for achieving high performance in time series forecasting. By leveraging the strengths of linear modeling, DLinear provides a robust and efficient alternative to more complex deep learning models. However, it has notable deficiencies, as it is ill-suited for modeling non-linear dependencies that may exist within the time-series sequence and between time-independent covariates. Consequently, for time-series forecasting tasks that require capturing intricate and non-linear interactions, more sophisticated models might be necessary to achieve optimal performance.

In this paper, we introduce a concise and efficient deep learning framework designed to significantly improve long-term predictions of atmospheric pollution concentrations compared to traditional neural network models. Our model, based on the Multi-Layer Perceptron (MLP), maintains simplicity by excluding complex self-attention or convolutional mechanisms. This design leads to linear computational growth as the context and forecast horizon lengths expand, making it more suitable for long-term predictions of pollutant concentrations compared to many Transformer-based models.

We propose the Wavelet Transform-based Multilayer Perceptron (WTMP) model architecture for long-term air pollution concentration forecasting. This model combines the signal decomposition algorithm of Maximal Overlap Discrete Wavelet Transform (MODWT) with an enhanced MLP structure. When applied to the air pollutant dataset from the Ruian Economic Development Zone in Zhejiang Province, our model demonstrates superior or similar performance compared to existing neural network-based benchmarks, achieving a Mean Squared Error reduction of over 15%. Additionally, WTMP is more than five times faster in both inference and training compared to the best Transformer-based model.

The effectiveness and feasibility of the proposed model were validated through a comprehensive set of experiments. Our findings reveal that the WTMP model has significant potential in accurately predicting long-term variations in atmospheric pollution concentrations, providing valuable insights for air quality monitoring and management. This innovative approach opens new avenues for future research in this field.

The remaining parts of this paper are arranged as follows. Section 2 discusses the relevant deep learning models. Section 3 presents the details of the proposed WTMP model. Section 4 provides the experiments and result analysis. Finally, Section 5 concludes the paper.

2. Background and Related Works

Models for long-term forecasting can be broadly categorized into multivariate and univariate models. Multivariate models leverage the historical data of multiple interrelated time series variables to predict their future behavior. Recent neural network-based multivariate models have gained traction, such as LongTrans [28], Informer [23], Autoformer [24], FEDFormer [25], and Pyraformer [26].

In contrast, univariate models rely solely on the historical data of a single time series and its covariates to predict its future. In this case, the historical data of other time series is not considered during prediction. Univariate models can be further divided into local and global models. Local univariate models are trained and inferred independently for each time series. Classical models such as Autoregressive (AR), Autoregressive Integrated Moving Average (ARIMA), exponential smoothing models, and the Box-Jenkins methodology fall into this category. Global univariate models use a shared model that is trained across the entire dataset and can be applied to different time series like DLinear model.

It is important to note that both multivariate and univariate models can be fairly compared in long-term forecasting tasks as long as they are evaluated on the same test set. In Section 4, we will follow this evaluation protocol to compare the performance of different models.

2.1. Residual Block

Throughout the model architecture, residual connection blocks are frequently employed (as shown in Figure 1). Each residual block is composed of a Multi-Layer Perceptron (MLP) with a hidden layer and employs ReLU activation functions to introduce non-linearity, thereby enhancing the model's capacity to capture complex features. A critical aspect of each residual block is its skip connection, which directly maps the input to the output, allowing gradients to flow uninterrupted through the network. This feature significantly improves the training efficiency of deep networks. Both the dropout and the layer norm can be turned on or off by a model level hyperparameter [29].

Our residual block design presents several advantages. Compared to architectures with complex self-attention mechanisms [20], our MLP-based structure is computationally more efficient, reducing both parameter count and computational complexity. The skip connection also alleviates the vanishing gradient problem, supporting the training of deeper networks. Furthermore, the incorporation of dropout in the skip connection's linear layer enhances the model's generalization and reduces overfitting.

Additionally, the design of the residual block allows for easy integration into various network architectures, making it versatile for different data types and tasks. Compared to residual blocks in other models [30], our design achieves a balance between simplicity and efficiency. Notably, in long-sequence time series forecasting tasks, our residual blocks have shown exceptional performance, as demonstrated by the experimental results.

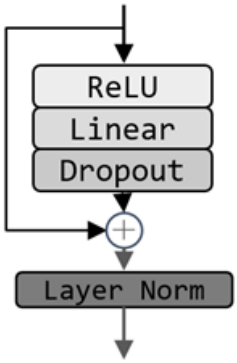


Figure 1. Residual Block.

2.2. Wavelet Decomposition

A wavelet is a wave-like oscillation localized in time, capable of capturing both the temporal and frequency characteristics of a signal [31]. This contrasts with the Fourier Transform, which solely captures the frequency domain. Wavelets can be categorized into two types: (a) Continuous Wavelet Transforms (CWT), which involve an infinite number of wavelets (i.e., every possible wavelet), and (b) Discrete Wavelet Transforms (DWT), which utilize a finite number of wavelets at specific times and locations. Our research focuses on the DWT, which represents a signal using an orthonormal basis representation. This method is widely employed in various fields, such as hydrology, epidemiology, and geophysics. The decomposition provided by DWT yields a set of time series, where each series has a coefficient describing the signal's temporal evolution within a specific frequency band [32]. The DWT operates on a discrete dyadic (octave) grid for the scale parameter i and shift parameter k. The equation for the forward DWT is given by [31]:

$$c(i,k) = \sum_t f(t)\Psi^*_{i,k}(t), \quad \text{where} \quad \Psi^*_{i,k}(t) = 2^{\frac{i}{2}}\Psi(2^i t - k). \tag{1}$$

The inverse DWT is defined as:

$$f(t) = \sum_k \sum_i c(i,k)\Psi_{i,k}(t), \tag{2}$$

where $f(\cdot)$ denotes the function and Ψ is the mother wavelet.

2.3. MODWT Algorithm

The Maximal Overlap Discrete Wavelet Transform (MODWT) is a variant of the Discrete Wavelet Transform (DWT) with several notable features that address the drawbacks and limitations of the DWT [33]. Both methods use a pair of filters to split the frequency range of the input data into low- and high-frequency components, referred to as scaling and wavelet coefficients, respectively. Unlike the DWT, the MODWT does not include a downsampling step in its decomposition process. As a result, it is generally interpreted that the computational load of the DWT is less than that of the MODWT. However, in matrix operations, all coefficients are calculated simultaneously, making it a single operation. For instance, considering an input signal with N samples, in the DWT, the $N/2$ wavelet coefficients are computed through the multiplication of an $N/2 \times N$ transformation matrix and an $N \times 1$ input signal. By contrast, the N wavelet coefficients of the MODWT are calculated using an $N \times N$ transformation matrix multiplied by an $N \times 1$ input signal.Additionally, in real-time applications, most of the processing time is spent on the computation of wavelet coefficients, which is significantly lower than the sampling frequency. Moreover, the MODWT offers advantages such as the ability to handle any sample size and signal shift invariance.

The Haar wavelet, favored for its computational simplicity, strong locality, and interpretability, stands out as the wavelet basis of choice for time series analysis [31]. While the db4, coif4, and fk4 wavelets may capture certain signal features more intricately, our ablation study in Section 4.3 revealed that their overall impact on atmospheric pollutant monitoring results is negligible, particularly with the coif4 wavelet performing similarly to the Haar wavelet. Given that the Haar wavelet is the simplest and most widely used option, offering superior computational efficiency, ease of use, and versatility, we concluded that it is the preferred choice for our monitoring and analysis efforts [34]. Additionally, the Haar wavelet's filter length of just 2 offers a significant advantage in terms of computational efficiency during the decomposition and reconstruction processes. Figure 2 is an example of using MODWT decomposition.

Below, we define the detail coefficients (denoted by $\tilde{s}_{j,l}$) and scaling coefficients (denoted by $\tilde{w}_{j,l}$) generated by an MODWT algorithm utilizing the Haar filter, which is among the most prevalent wavelet filters in time series literature.

$$\tilde{d}_{j,l} = \frac{d_{j,l}}{2^{j/2}}, \quad \tilde{s}_{j,l} = \frac{s_{j,l}}{2^{j/2}} \tag{3}$$

The wavelet and scaling filters, denoted as $d_{(j,l)}$ and $s_{(j,l)}$, respectively, convolve the coefficients of the j-th level (where $l = 0, 1, \ldots, L-1$) from the original time series $\{Y_t : t = 0, 1, \ldots, N-1\}$ to generate the detailed coefficients ($D_{(\tilde{j},t)}$) and scaling coefficients ($S_{(\tilde{j},t)}$) [31] as follows:

$$\tilde{D}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{d}_{j,l} Y_{(t-l)modN}, \quad \tilde{S}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{s}_{j,l} Y_{(t-l)modN} \tag{4}$$

where ($D_{(\tilde{j},t)}$). The length of the decomposition into different scale detail coefficients is consistent with the length of the original signal.

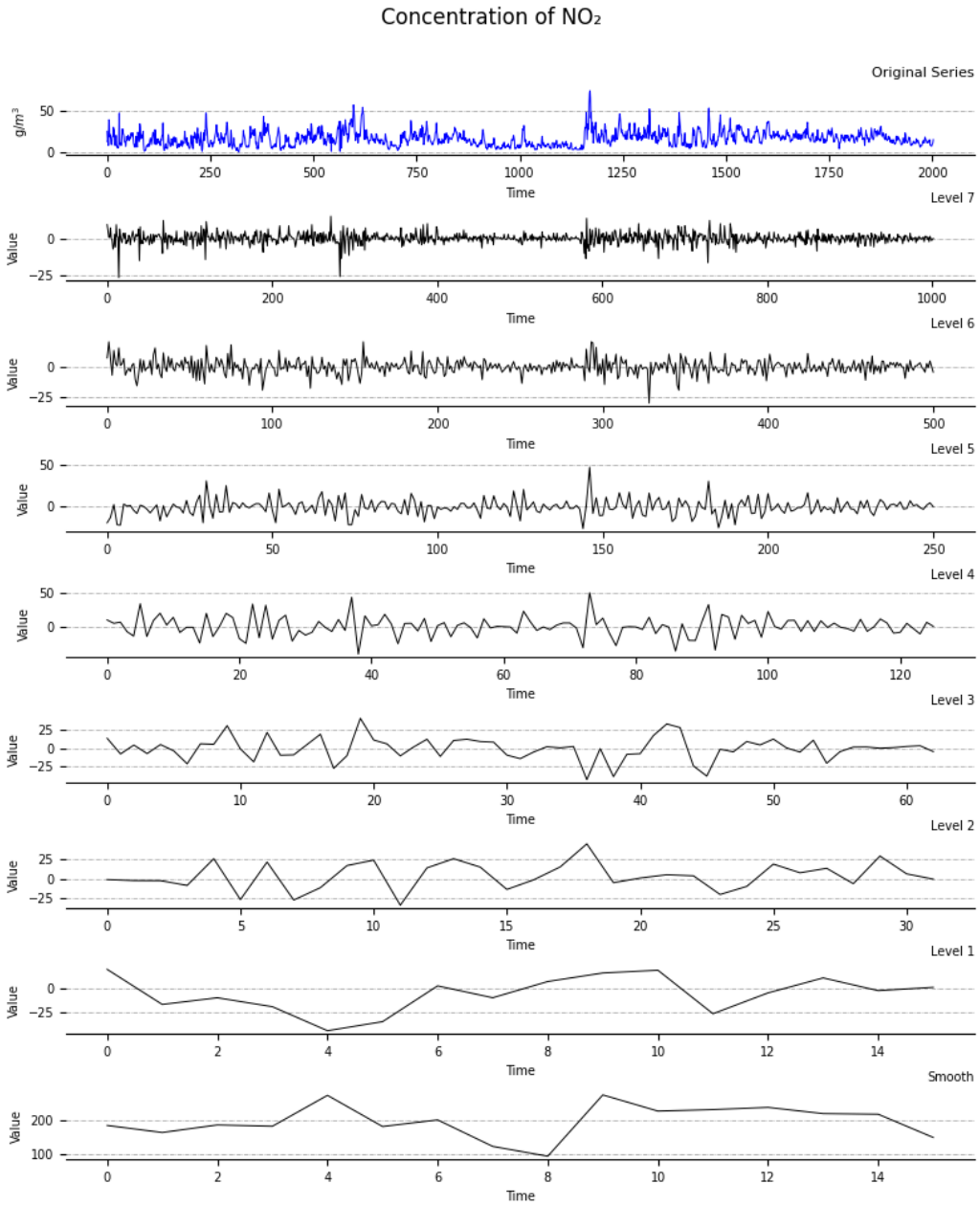


Figure 2. MODWT decomposition of the NO_2 dataset. The first curve represents the original time series; the last represents the smooth series, i.e., the scaling coefficient of the time series after computing the MODWT algorithm. The other seven curves (in between) represent the different level coefficient calculated by the MODWT algorithm. The filter used for this decomposition is the Haar filter.

2.4. Decomposition Level Selection

In time series forecasting, selecting the number of MODWT decomposition levels is a critical step that determines the extent of frequency components that can be captured. Setting the number of decomposition levels $J+1$ to $\lfloor \log_e N \rfloor$ is a method based on both theory and experience [31]. It ensures that the number of decomposition levels is proportional to the length of the time series N, preventing an excess that could increase noise or a deficiency that might miss vital information. This approach provides the capability for multi-scale analysis, allowing the model to extract features from different temporal scales, thereby better understanding and forecasting the behavior of the time series. Additionally, this setting is flexible and can adapt to time series data of varying lengths, achieving a good

balance between theory and practice, and supporting the enhancement of the predictive model's performance.

2.5. Covariate Information

Covariates are commonly used in statistics and data analysis to refer to variables that may influence the outcome of a primary dependent variable. While these covariates are not the main focus of the study, they may be associated with the primary variable and therefore need to be accounted for and controlled in the analysis. Covariates can be categorized into static and dynamic types [29]. Dynamic covariates change over time, such as wind speed, wind direction, or temperature. In contrast, static covariates are variables that do not change or can be determined in advance, such as the date.

We use $x_i^t \in \mathbb{R}^r$ to denote the r-dimensional covariates of time-series i at time t. These covariates can include variables specific to the time-series, such as the day of the week and holiday information. Although these variables vary over time, they are known in advance and are thus more appropriately treated as static covariates. In many applications, incorporating such covariates has been shown to significantly improve forecasting accuracy, and a well-designed model architecture should effectively integrate them.

2.6. Problem Setting

Let $i : j$ denote the data set $\{i, i+1, \ldots, j\}$. We define the input of pollutant data at time t as $y^t_{(1:L)}$, and $x^t_{(1:L+H)}$ as the input time data at time t, which includes both known sequences and forecasting sequence time information. The predicted pollutant results at time t are represented by $\tilde{y}^t_{(L+1:L+H)}$. Here, L and H represent the lengths of the input and output sequences, respectively. The dimension of x is r, and the feature dimensions of y and $\tilde{y}$ are denoted by d_y and $d_{\tilde{y}}$, respectively, representing the number of pollutant species for input and output.Let the predictive model function be denoted as f.

The prediction process is then defined as:

$$f : \left(\{y^t_{(1:L)}\}_{t=1}^N, \{x^t_{(1:L+H)}\}_{t=1}^N \right) \rightarrow \{\tilde{y}^t_{(L+1:L+H)}\}_{t=1}^N. \tag{5}$$

3. Methodology

In the current section, we provide an in-depth description of our model. With reference to Figure 3, we propose an enhanced MLP network model that synergizes historical data with both time series and wavelet characteristics to modify the DLinear model. To address the inherent discontinuities in the temporal monitoring of atmospheric pollutant concentrations, we leverage a sophisticated GAN [35] for effective data interpolation. Thereafter, the refined dataset is integrated into our model framework, which employs a modwt decomposition, enriched with temporal covariate data, to yield the forecasted results.

3.1. Data Description

The data is sourced from the environmental monitoring station located at coordinates 27.778203, 120.655098 in the Economic Development Zone of Ruian City. Ruian Economic Development Zone is located in the southeast coast of Zhejiang Province, with a terrain dominated by hills and mountains. The land is higher in the west and lower in the east, with the western part being a region of medium and low hills and mountains, generally above 600 m in altitude. The industry in the area is well-developed, forming industries including automotive parts, polymer materials, and mechanical electronics. The station monitors various atmospheric pollutants including nitrogen oxides, sulfur oxides, carbon oxides, ozone, PM10, PM2.5, as well as meteorological indicators such as wind speed, wind direction, temperature, and relative humidity. Data indicators with missing values accounting for more than 10% of the total samples are excluded, resulting in a dataset comprising 104 data indicators. Both the original and prediction sequences in the experiments are set to have dimensions of 104. The experimental dataset consists of historical

monitoring data from 1 January 2021 to 30 June 2022, with a sampling interval of 1 h, totaling 13,095 data points.

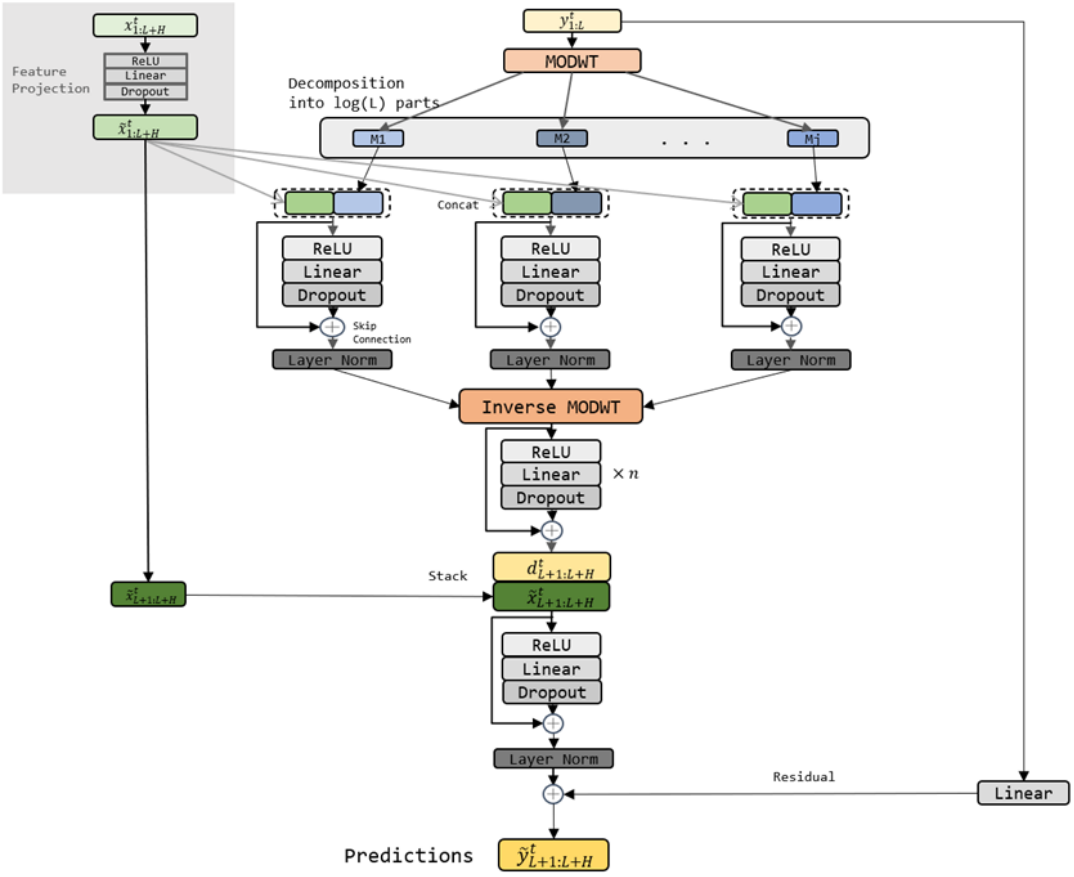


Figure 3. The architecture of the WTMP model.

The data is sourced from the environmental monitoring station in the Economic Development Zone of Ruian City. The station monitors various atmospheric pollutants including nitrogen oxides, sulfur oxides, carbon oxides, ozone, PM10, PM2.5, as well as meteorological indicators such as wind speed, wind direction, temperature, and relative humidity. During the data preprocessing phase, the collected raw data went through multiple steps. Initially, data entries with more than ten percent missing values were removed, followed by the elimination of irrelevant attributes. Table 1 provides an overview of the attributes and their corresponding symbols in each dataset after the completion of the preprocessing stage.

Table 1. Attribute Comparison Table.

Name	Unit	Name	Unit
Time	year-month-day-hour	SO_2	$\mu g/m^3$
NO_2	$\mu g/m^3$	O_3	$\mu g/m^3$
$PM_{2.5}$	$\mu g/m^3$	PM_{10}	$\mu g/m^3$
Wind speed	m/s	Wind direction	°
Temperature	°C	Isobutane	ppb
Ethane	ppb	Propane	ppb
Butane	ppb	Isopentane	ppb
2,3-Dimethylbutane	ppb	2-Methylpentane	ppb

Table 1. *Cont.*

Name	Unit	Name	Unit
3-Methylpentane	ppb	n-Hexane	ppb
Methylcyclopentane	ppb	2,4-Dimethylpentane	ppb
Cyclohexane	ppb	2-Methylhexane	ppb
2,3-Dimethylpentane	ppb	3-Methylhexane	ppb
2,2,4-Trimethylpentane	ppb	n-Heptane	ppb
Methylcyclohexane	ppb	2,3,4-Trimethylpentane	ppb
2-Methylheptane	ppb	3-Methylheptane	ppb
n-Octane	ppb	n-Nonane	ppb
n-Decane	ppb	n-Undecane	ppb
n-Dodecane	ppb	1-Hexene	ppb
Ethylene	ppb	Propylene	ppb

3.2. Data Processing

In the actual training phase of the WTMP model, we need time series data that are as continuous as possible, and each record must have complete annotations. However, due to the missing data in the time series, we need to adopt appropriate methods to initialize the data and obtain the covariate information within the data. The data processing methods in this paper include data imputation.

Accurately imputing missing values in datasets is critical for subsequent analyses. To address this challenge, Yoon et al. introduced the Generative Adversarial Imputation Network (GAIN) [35]. This model leverages the framework of Generative Adversarial Networks (GANs), where a generator and a discriminator are trained adversarially to produce plausible imputations for missing data.

The GAIN model's principal virtue is its adeptness at capitalizing on the comprehensive data distribution to deduce missing values, yielding more authentic and credible imputations. As delineated in Figure 4, GAIN distinguishes itself from conventional imputation strategies—like mean and regression imputation—and even from sophisticated methodologies such as MissForest [36] and Autoencoders [37], by adeptly leveraging the global data distribution. This proficiency enables GAIN to craft imputations that transcend mere reliance on localized correlations, instead embracing the holistic attributes of the dataset.

Our extensive array of experiments, conducted across a tapestry of datasets with missing data ranging from a mere 2% to a substantial 20%, consistently underscored the GAIN algorithm's superior efficacy across the continuum of data incompleteness. This distinctive edge endows GAIN with a unique advantage for multidimensional data applications, particularly excelling in the domain of atmospheric pollutant forecasting.

Our experiments with atmospheric pollutant datasets not only confirm the superior performance of the GAIN algorithm but also highlight its adaptability and precision in environmental data analysis. These features make GAIN an essential tool for environmental scientists and policymakers, playing a significant role in developing effective strategies to address atmospheric pollution.

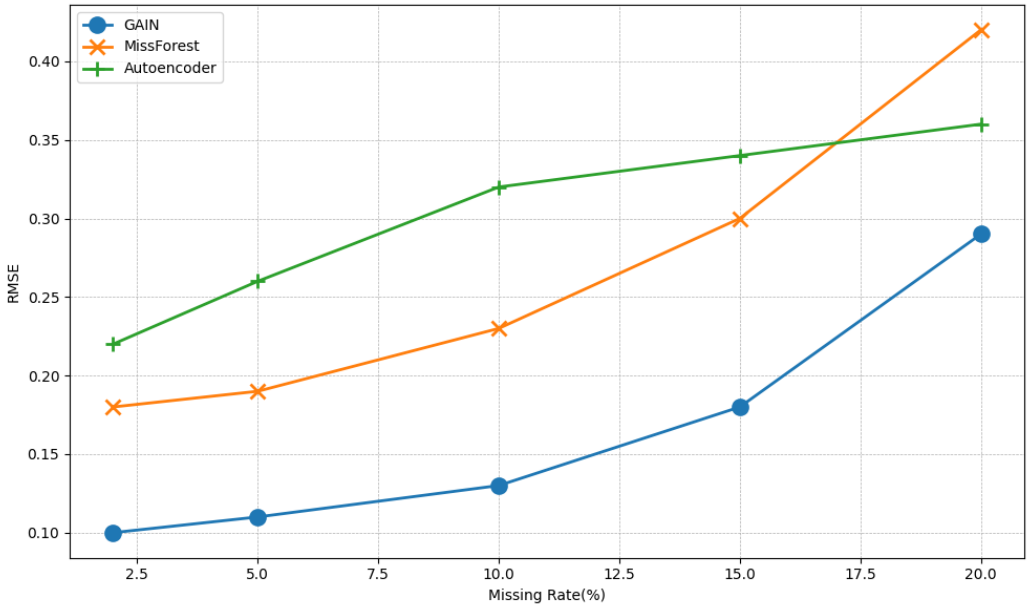


Figure 4. RMSE performance in various missing rates.

3.3. Model Architecture

For the input pollutant raw data $y^t_{1:L}$, it is initially decomposed via the MODWT to capture more explicit information about the observed time series. ethm will provide us with J higher-frequency detail coefficients and a smooth (trend component) of the available time series. In Figure 2, we present a visual representation of the MODWT decomposition applied to the NO_2 concentration series in atmospheric pollution data, utilizing a Haar filter. The MODWT decomposition can be represented as:

$$F_{\text{modwt}}(y^t_{1:L}) = \sum_{j=1}^{J} M_j + S \tag{6}$$

Here, M_j represents the j-th level details, and S is the smooth component of the decomposed time series.The MRA-based MODWT decomposed series can handle nonstationarity and seasonality in the time series datasets and capture the overall trend of the series.

As depicted in the top-left of Figure 3, for the input static covariates $x^t_{1:L+H}$, which include information such as time, date, holidays, known in advance for both the input sequence (of length L) and the prediction sequence (of length H) [38], we utilize a residual block to map the dimensionality from r to $\tilde{r}$ ($\tilde{r} \ll r$) , resulting in the mapped sequence $\tilde{x}^t_{1:L+H}$:

$$\tilde{x}^t_{1:L+H} = \text{ResidualBlock}(x^t_{1:L+H}) \tag{7}$$

This essentially performs dimensionality reduction, preventing the flattening of static covariates into a large one-dimensional vector of size $(L+H)r$. This reduction diminishes the weight of pollutant data during training. After dimensionality reduction, the flattened vector size is only $(L+H)\tilde{r}$.

For the sequence of covariates that has been dimensionality reduced, $\tilde{x}^t_{(1:L+H)}$, in order to incorporate covariate information into the sequence, $\tilde{x}^t_{(1:L+H)}$ is concatenated with the wavelet-decomposed M_j and S. The concatenation method involves flattening the two sequences into a one-dimensional connection, resulting in $J+1$ vectors of size $(L+H)\tilde{r} + L \cdot d_y$. Assuming the output of the residual block is denoted as h, then through the residual block, we obtain vectors $\tilde{M}_j$ and $\tilde{S}$ of length h:

$$\tilde{M}_j = \text{ResidualBlock}(M_j; \tilde{x}^t_{1:L}), \quad j = 0, 1, \ldots, J \tag{8}$$

$$\tilde{S} = \text{ResidualBlock}(S; \tilde{x}^t_{1:L}) \tag{9}$$

Next, we perform an inverse wavelet transform on $\tilde{M}_j$ and $\tilde{S}$ to obtain one-dimensional vectors of length h. These vectors are then passed through n residual blocks to generate a one-dimensional prediction component g^t of length $H \cdot p$, where H represents the output sequence length, p denotes the output dimension, and n is the number of residual blocks, a hyperparameter.

$$g^t = \text{ResidualBlock}\left(F_{\text{imodwt}}(\tilde{M}_1, \ldots, \tilde{M}_J, \tilde{S})\right) \in \mathbb{R}^{p \cdot H} \tag{10}$$

We reshape g^t to obtain the prediction component $d^t_{(L+1):(L+H)}$ in $\mathbb{R}^{H \times p}$:

$$d^t_{(L+1):(L+H)} = \text{Reshape}(g^t) \in \mathbb{R}^{H \times p} \tag{11}$$

Then, we concatenate the prediction component $d^t_{(L+1):(L+H)}$ with the covariates $\tilde{x}^t_{(L+1):(L+H)}$, projecting the transformed covariates onto the prediction component. This results in a vector of dimension $H \times (p + \tilde{r})$, which is passed through another residual block to generate the predicted result $\tilde{d}^t_{(L+1):(L+H)}$:

$$\tilde{d}^t_{(L+1):(L+H)} = \text{ResidualBlock}\left(d^t_{(L+1):(L+H)}; \tilde{x}^t_{(L+1):(L+H)}\right) \in \mathbb{R}^{H \times \tilde{\delta}_y} \tag{12}$$

To strengthen the influence of covariates, particularly for prediction tasks like atmospheric pollutant prediction where date and holiday features have an impact on actual values, we reintroduce the covariates of the future prediction sequence onto the prediction component. This inclusion of covariates acts as a "highway" to prevent the loss of covariate information.

Finally, we add a global residual connection layer within the model [29]. Through a linear layer, the original data $y^t_{1:L}$ is mapped, and then added to $\tilde{d}^t_{(L+1):(L+H)}$ to obtain the final prediction result $\tilde{y}^t_{(L+1):(L+H)}$. This linear layer essentially represents the core of the DLinear model. By incorporating a global residual connection, the WTMP model theoretically performs no worse than the DLinear model. In mathematical notation:

$$\tilde{y}^t_{(L+1):(L+H)} = y^t_{(L+1):(L+H)} + \text{LinearLayer}(y^t_{1:L}) \tag{13}$$

3.4. Evaluation

The precision of the predictive model's forecasts is gauged by a metric that assesses the proximity of its predictions to the true outcomes. Specifically, Mean Squared Error (MSE) is utilized as the training loss function, the lower values of MSE indicate better forecasting results.the quality of the model's fit is evaluated by the following criterion:

$$MSE\left(\left\{y^t_{(L+1:L+H)}\right\}_{t=1}^{N}, \left\{\tilde{y}^t_{(L+1:L+H)}\right\}_{t=1}^{N}\right) = \frac{1}{NH}\sum_{t=1}^{N}\left\|y^t_{(L+1:L+H)} - \tilde{y}^i_{(L+1:L+H)}\right\|_2^2. \tag{14}$$

An epoch is comprised of all possible look-back and forecast pairs that can be formed from the training data, allowing for potential overlap between mini-batches. This approach aligns with standard practices in previous studies on long-term forecasting.

Evaluation of the model is conducted on a test set, examining every possible look-back and forecast pair that can be derived from the test data. This method is commonly referred to as rolling validation or evaluation. Optionally, a similar evaluation can be applied to a validation set to fine-tune model parameters for optimal selection.

4. Experiments

In our research, we utilize Min–Max normalization to bring the dataset into a standard form. Subsequently, the data is partitioned into three distinct subsets: training, valida-

tion, and testing, with the proportions being 70%, 20%, and 10%, respectively. For the training phase, we implement a batch size of 64 and leverage the Adam optimization algorithm, which is set with a learning rate of 0.001. The training process is designed to halt after 200 epochs or if there is no reduction in the loss for 15 successive iterations. The reported metrics are derived from the mean of five independent experimental runs to ensure reliability and consistency.

4.1. Baselines and Setup

We choose SOTA Tranformer based models for time-series including Transformer, Informer, Pyraformer, Autoformer. Recent findings by DLinear have demonstrated that straightforward linear models can surpass the above models. Consequently, the DLinear model stands as a critical benchmark for comparison.Furthermore, we have decided to include an LSTM in our comparative evaluation [39], utilizing a one-layer LSTM with a 128-dimensional configuration.

In this study, we aim to accurately predict atmospheric pollutant concentrations in a specific region. Following widely adopted benchmarking standards, we selected 336 and 720 h as the look-back windows and {96, 168, 336} h as the forecast horizons. These time spans are chosen to capture mid-term trends and seasonal patterns in the time series, with data sampled on an hourly basis. The two- to four-week observation period not only covers short-term fluctuations but also accounts for potential cyclical factors. The forecast horizons are set to 4 to 14 days, providing an effective range for distinguishing the long-term prediction performance of different models while avoiding excessive errors that can arise from overly extended forecasts, which might obscure the true differences in model performance.

4.2. Results

The Mean Squared Error (MSE) and Mean Absolute Error (MAE) for all models at different forecast lengths are presented in Table 2. The final results are the average values taken from 5 independent runs, with the best experimental results in bold for the same forecast length.

Table 2. Comparison of Different Prediction Horizons of Atmospheric Pollutant Data.

Models	336/96	336/168	720/336	720/168
Transformer	0.698 0.487	0.674 0.567	0.691 0.475	0.655 0.548
Pyraformer	0.580 0.563	0.582 0.511	0.865 0.423	0.635 0.496
Autoformer	1.190 0.830	4.657 1.505	1.811 0.532	1.137 1.639
Informer	0.643 0.576	0.639 0.545	0.654 0.555	0.645 0.574
LSTM	0.611 0.554	0.720 0.652	0.812 0.651	0.812 0.598
DLinear	0.482 0.442	0.517 0.442	0.706 0.477	0.679 0.479
WTMP	**0.341 0.377**	**0.387 0.412**	**0.545 0.415**	**0.526 0.417**

Note: The entry "336/96" indicates that the original data sequence is 336 points long, with a prediction sequence of 96 points; subsequent entries follow the same format.

Furthermore, we have applied our trained models to real-world data to evaluate and compare their predictive capabilities, with the comparative curves illustrated in the Figure 5. Our findings indicate that initially, the LSTM model's forecasting accuracy is competitive with those based on Transformer architectures. Nonetheless, with an extended forecast sequence, there is a gradual increase in cumulative errors, which diminishes its predictive accuracy. The DLinear model also shows superiority over Transformer-based models for shorter forecast sequences but faces challenges in capturing the complex dynamics between pollutants as the forecasting horizon grows. Even so, it retains a level of performance that is still respectable. The WTMP model, enhancing the foundation laid by DLinear, demonstrates the most impressive practical performance, thanks to the MODWT decomposition that adeptly manages non-stationary data characteristics. By sidestepping the complexity of the self-attention mechanism, the MLP within the WTMP model excels

particularly in long-term forecasting tasks. The experimental outcomes underscore its markedly superior forecasting performance in comparison to other models.

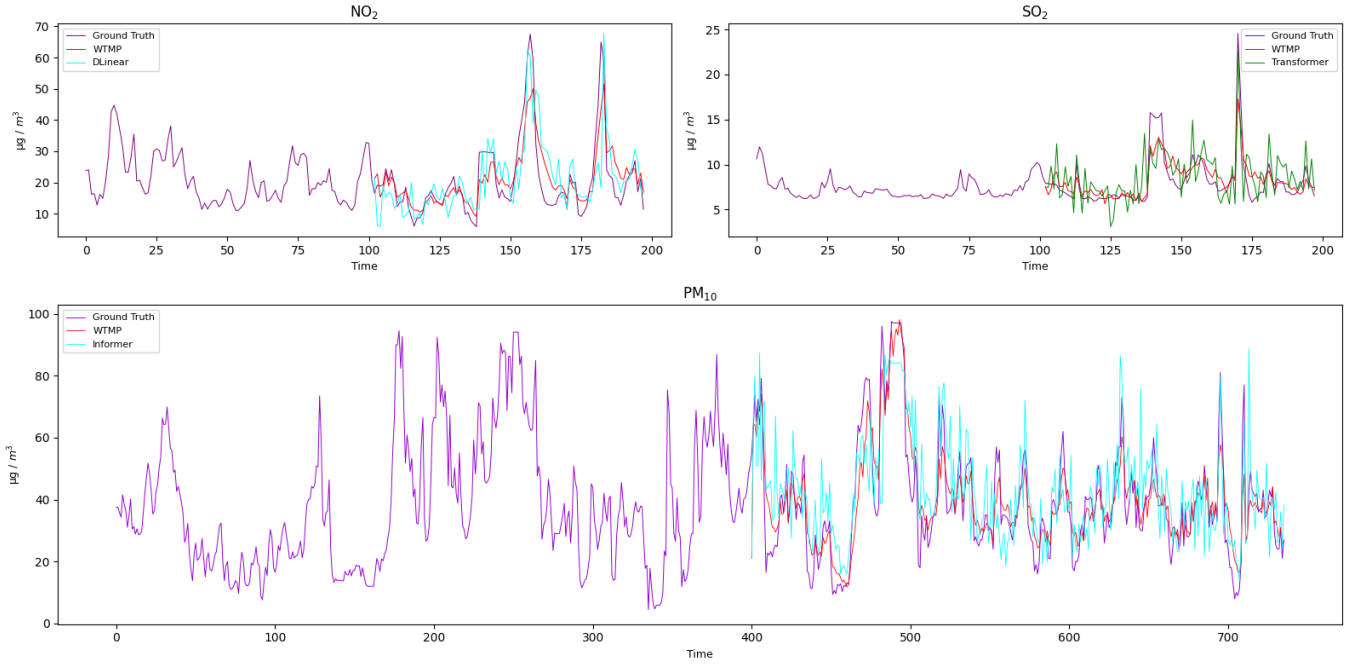


Figure 5. Trained model used for predicting different pollutant data and generating long-term forecast curves.

4.3. Ablation Study

We aimed to validate the efficacy of the MODWT algorithm, particularly when utilizing the Haar wavelet basis, in the context of time series forecasting for atmospheric pollutant concentrations. To this end, we devised an ablation study comparing various models with distinct configurations of wavelet transformations and other time series decomposition techniques.

4.3.1. Model A: Direct Feature Fusion

This model is a variation of the WTMP framework, where we have intentionally omitted the wavelet transformation process. Instead, the model directly integrates the features from residual blocks with covariate information after multiple passes, yielding the final output. This setup serves as a control to assess the impact of wavelet transformation on predictive performance.

4.3.2. Model B: Sliding Average Decomposition

Inspired by the effectiveness demonstrated by the sliding window average method in the DLinear model, we replaced the MODWT algorithm with a simple moving average technique in this model. The decomposed input sequence is merged by a straightforward addition, emphasizing the simplicity and efficiency of this approach. This model tests the hypothesis that a more traditional smoothing technique can be competitive with the MODWT algorithm.

4.3.3. Model C: MODWT with db4 Wavelet Basis

Here, we selected the 'db4' wavelet basis for the MODWT algorithm to explore the impact of wavelet choice on the forecasting accuracy. The 'db4' basis is known for its smoothness properties, which may be beneficial for certain types of time series data [40].

4.3.4. Model D: MODWT with coif4 Wavelet Basis

In this model, we opted for the 'coif4' wavelet basis, which is designed to provide a more detailed analysis of the time series, potentially capturing higher-frequency variations that could be significant in pollution data.

4.3.5. Model E: MODWT with fk4 Wavelet Basis

We chose the 'fk4' wavelet basis for this model, which is known for its balance between smoothness and detail, offering a comprehensive representation of the time series at various scales.

As before, we processed the validation dataset through the meticulously trained models. By scrutinizing the Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics, we discerned the comparative efficacy of each model. The Table 3 presents the empirical results of our experiment.

Table 3. Comparison of Different Prediction Horizons of Atmospheric Pollutant Data.

Models	336/96	336/168	720/336	720/168
WTMP	0.341 0.377	0.387 0.412	0.545 0.415	0.526 0.417
Model A (Direct Feature Fusion)	0.763 0.901	0.892 0.628	1.163 0.837	1.121 0.831
Model B (Sliding Average Decomposition)	0.631 0.510	0.692 0.561	0.796 0.640	0.764 0.603
Model C (db4)	0.410 0.386	0.423 0.465	0.596 0.458	0.586 0.429
Model D (coif4)	0.333 0.370	0.381 0.422	0.552 0.422	0.525 0.423
Model E (fk4)	0.461 0.390	0.399 0.409	0.610 0.430	0.596 0.421

From the experimental outcomes, we deduce the following conclusions: Model A, devoid of any decomposition technique, exhibits the weakest predictive performance, with a notably increased disparity from WTMP as the forecast horizon lengthens. Model B, employing a sliding average method, is simpler than wavelet transformations but potentially lacks the finesse to capture the intricate temporal features that MODWT can discern, thus demonstrating subpar predictive capabilities. Subsequently, we have Models C, D, and E that incorporate MODWT decomposition. Model D, leveraging the coif4 wavelet basis, is particularly adept at capturing high-frequency fluctuations and thus rivals the performance of Haar, occasionally outperforming it. Conversely, Models C and E, which employ the 'db4' and 'fk4' wavelet bases respectively, have demonstrated inferior performance in the experiment. They are not adept at managing the variables involved in this study.

4.4. More Analyses on Different Models

The LSTM model employs gating mechanisms, but the sequential processing of sequences can lead to error accumulation and vanishing gradients. In contrast, the Transformer model, which uses self-attention mechanisms, can more effectively capture global dependencies, thus compensating for the vanishing gradient problem. The improved Transformer architecture, Informer, utilizes the DMS (Direct Multi-Step) forecasting strategy to address error accumulation. However, the fundamental nature of the Transformer model is to model the semantic association between paired elements, thereby disrupting the temporal order of the elements. Although the introduction of positional encoding and the use of tokens to embed subsequences can help retain some sequential information, this is fatal for time series forecasting. The Dlinear model points out this shortcoming, retaining the sequential characteristics of time series and achieving results close to or surpassing those of Transformer models with just two simple linear layers. However, due to its simplicity, it abandons further processing of temporal information, making it difficult to model non-linear dependencies among the time-series sequence and time-independent covariates. The WTMP model designed based on this effectively addresses the aforementioned issues.

5. Conclusions

We propose a novel forecasting model known as WTMP, which integrates wavelet decomposition into an enhanced MLP model. Predictive experiments conducted using atmospheric pollutant data from the Rui'an Economic Development Zone have demonstrated the effectiveness of this method for long-sequence forecasting tasks of industrial park air pollutant concentrations. The key findings are as follows:

1. Prediction models based on the Transformer architecture underperform when compared to a simple MLP model. At least for the task of atmospheric pollutant forecasting, there is no need to employ the complex self-attention mechanism—central to Transformer models—to learn the periodicity and trends of long sequences.
2. The introduction of a wavelet decomposition module and residual connection blocks within the model has significantly improved the capture of non-stationarity and non-linear dependencies within the sequences. This enhancement has further elevated the predictive performance and effectively circumvented the issues of low fitting accuracy and poor generalization associated with linear models.

6. Future Work

Next, we will investigate the limitations of the WTMP model and verify its generalization ability by testing it on pollutant data from other regions. Where feasible, we will develop an atmospheric pollutant concentration model based on spatiotemporal data from multiple monitoring stations.

Author Contributions: Conceptualization, L.T., M.F. and X.W.; Methodology, L.T. and M.F.; Software, M.F.; Validation, L.T.; Formal Analysis, L.T.; Investigation, Q.W.; Resources, Q.W.; Data Curation, Q.W.; Writing—Original Draft Preparation, L.T.; Writing—Review and Editing, X.W.; Visualization, M.F.; Supervision, X.W.; Project Administration, X.W.; Funding Acquisition, X.W. and M.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62033012).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study will be available from the corresponding author upon reasonable request.

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

References

1. Chanchan, G.; Fengying, Z.; Dekun, F.; Qingtao, W.; Min, L. Spatial characteristics of change trends of air pollutants in Chinese urban areas during 2016–2020: The impact of air pollution controls and the COVID-19 pandemic. *Atmos. Res.* **2023**, *283*, 106539.
2. Dw, D. An association between air pollution and mortality in six US cities. *N. Engl. J. Med.* **1993**, *329*, 1753–1759.
3. Heft-Neal, S.; Burney, J.; Bendavid, E.; Burke, M. Robust relationship between air quality and infant mortality in Africa. *Nature* **2018**, *559*, 254–258. [CrossRef]
4. Weijun, W.; Tianyu, M.; Lianru, W. Air pollutant concentration prediction based on a new hybrid model, feature selection, and secondary decomposition. *Air Qual. Atmos. Health* **2023**, *16*, 2019–2033.
5. Chen, L.; Han, B.; Wang, X.; Zhao, J.; Yang, W.; Yang, Z. Machine learning methods in weather and climate applications: A survey. *Appl. Sci.* **2023**, *13*, 12019. [CrossRef]
6. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
7. Arystanbekova, N.K. Application of Gaussian plume models for air pollution simulation at instantaneous emissions. *Math. Comput. Simul.* **2004**, *67*, 451–458. [CrossRef]
8. Deb, C.; Zhang, F.; Yang, J.; Lee, S.E.; Shah, K.W. A review on time series forecasting techniques for building energy consumption. *Renew. Sustain. Energy Rev.* **2017**, *74*, 902–924. [CrossRef]
9. Tao, Q.; Liu, F.; Li, Y.; Sidorov, D.N. Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU. *IEEE Access* **2019**, *7*, 76690–76698. [CrossRef]
10. David, S.; Valentin, F.; Jan, G.; Tim, J. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **2020**, *36*, 1181–1191.

11. Hoedt, P.J.; Kratzert, F.; Klotz, D.; Halmich, C.; Holzleitner, M.; Nearing, G.S.; Hochreiter, S.; Klambauer, G. Mc-lstm: Mass-conserving lstm. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 4275–4286.
12. Awasthi, P.; Das, A.; Sen, R.; Suresh, A.T. On the benefits of maximum likelihood estimation for regression and forecasting. *arXiv* **2021**, arXiv:2106.10370.
13. Shumway, R.H.; Stoffer, D.S.; Shumway, R.H.; Stoffer, D.S. ARIMA models. In *Time Series Analysis and Its Applications: With R Examples*; Springer: New York, NY, USA, 2017; pp. 75–163.
14. Dhiman, H.S.; Deb, D.; Guerrero, J.M. Hybrid machine intelligent SVR variants for wind forecasting and ramp events. *Renew. Sustain. Energy Rev.* **2019**, *108*, 369–379. [CrossRef]
15. Wu, Q.; Lin, H. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Sci. Total Environ.* **2019**, *683*, 808–821. [CrossRef] [PubMed]
16. Wu, C.; Li, Q.; Hou, J.; Karimian, H.; Chen, G. PM2.5 concentration prediction using convolutional neural networks. *Sci. Surv. Mapp* **2018**, *43*, 68–75.
17. Huang, C.J.; Kuo, P.H. A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. *Sensors* **2018**, *18*, 2220. [CrossRef] [PubMed]
18. Pak, U.; Kim, C.; Ryu, U.; Sok, K.; Pak, S. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. *Air Qual. Atmos. Health* **2018**, *11*, 883–895. [CrossRef]
19. Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A hybrid model for spatiotemporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory. *Sci. Total Environ.* **2019**, *664*, 1–10. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
21. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
22. Lee, J.; Toutanova, K. Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
23. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 11106–11115.
24. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
25. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 27268–27286.
26. Liu, S.; Yu, H.; Liao, C.; Li, J.; Lin, W.; Liu, A.X.; Dustdar, S. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
27. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 11121–11128. [CrossRef]
28. Lyu, H.; Sha, N.; Qin, S.; Yan, M.; Xie, Y.; Wang, R. Advances in neural information processing systems. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
29. Das, A.; Kong, W.; Leach, A.; Sen, R.; Yu, R. Long-term forecasting with tide: Time-series dense encoder. *arXiv* **2023**, arXiv:2304.08424.
30. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
31. Rhif, M.; Ben Abbes, A.; Farah, I.R.; Martínez, B.; Sang, Y. Wavelet transform application for/in non-stationary time-series analysis: A review. *Appl. Sci.* **2019**, *9*, 1345. [CrossRef]
32. Yoo, Y.; Baek, J.G. A novel image feature for the remaining useful lifetime prediction of bearings based on continuous wavelet transform and convolutional neural network. *Appl. Sci.* **2018**, *8*, 1102. [CrossRef]
33. Sasal, L.; Chakraborty, T.; Hadid, A. W-transformers: A wavelet-based transformer framework for univariate time series forecasting. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; pp. 671–676.
34. Heddam, S.; Merabet, K.; Difi, S.; Kim, S.; Ptak, M.; Sojka, M.; Zounemat-Kermani, M.; Kisi, O. River water temperature prediction using hybrid machine learning coupled signal decomposition: EWT versus MODWT. *Ecol. Inform.* **2023**, *78*, 102376. [CrossRef]
35. Yoon, J.; Jordon, J.; Schaar, M. Gain: Missing data imputation using generative adversarial nets. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 5689–5698.
36. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
37. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 353–374.
38. Lara-Benítez, P.; Carranza-García, M.; Luna-Romera, J.M.; Riquelme, J.C. Temporal convolutional networks applied to energy-related time series forecasting. *Appl. Sci.* **2020**, *10*, 2322. [CrossRef]

39. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
40. Safara, F.; Doraisamy, S.; Azman, A.; Jantan, A.; Ramaiah, A.R.A. Multi-level basis selection of wavelet packet decomposition tree for heart sound classification. *Comput. Biol. Med.* **2013**, *43*, 1407–1414. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.