Journal of Computer and Knowledge Engineering, Vol. , No.. 2023.
DOI:

Overcoming the Persistence Barrier in Air Quality Forecasting: A Zero-Initialized Residual Bi-GRU Framework*

First A. Author[1], Second B. Author[2], Jr., and Third C. Author[3]

***Abstract*--Accurate short-term $PM_{2.5}$ forecasting is vital for public health, yet remains difficult due to the stochastic nature of pollutants. Traditional neural networks often struggle with this task, frequently defaulting to simplistic identity transformations when processing data with strong temporal dependencies. To address these limitations, this paper introduces the Zero-Initialized Residual Bidirectional GRU (ZIR-BiGRU) framework.**

Unlike standard models that predict absolute pollution levels, the ZIR-BiGRU focuses on learning non-linear residuals—deviations from the current state—combined with a specialized zero-initialization strategy that ensures training stability from a persistence-equivalent baseline. Evaluated against real-world hourly air quality data, the model demonstrates robust predictive performance across a 24-hour horizon. Experimental results show significant improvements over the persistence baseline, achieving a 20.60% increase in the Coefficient of Determination (R^2) and a 4.37% reduction in Root Mean Squared Error (RMSE), with statistical significance confirmed by the Diebold-Mariano test ($p < 0.05$). Horizon analysis reveals that the model excels after a critical crossover point of $h = 6$, demonstrating a superior ability to capture complex long-term dependencies and sudden fluctuations in air quality compared to existing approaches.

***Index Terms*-- Air Quality Forecasting; Bidirectional GRU; Deep Residual Learning; Prediction; Time Series Analysis; Zero-Initialization**

1. Introduction

Air pollution, particularly fine particulate matter with an aerodynamic diameter of less than 2.5 micrometers ($PM_{2.5}$), has emerged as one of the most critical global health challenges of the 21st century. According to the World Health Organization (WHO), long-term exposure to elevated $PM_{2.5}$ concentrations is strongly correlated with severe respiratory diseases, cardiovascular dysfunctions, and increased mortality rates [1]. Beyond public health, particulate pollution significantly degrades atmospheric visibility and imposes substantial economic burdens on urban centers [2]. Consequently, establishing an accurate, robust, and early-warning forecasting system for $PM_{2.5}$ is not merely an academic pursuit but a societal imperative for effective environmental governance and public health protection.

In recent years, the development of deep learning models has brought about a significant transformation in air quality prediction. Recent research has shown that models such as LSTM, GRU, bi-directional networks (Bi-LSTM and Bi-GRU), and attention-based architectures are able to extract complex temporal dependencies and nonlinear relationships in air pollutant data with greater accuracy. In addition, the use of residual learning strategies and hybrid architectures has improved model convergence, increased stability of the training process, and improved prediction performance, especially at longer time horizons. These advances indicate that the development of dedicated deep learning architectures for air quality data remains one of the most important research topics in this field [3,6].

Forecasting air quality is an inherently complex task due to the stochastic and non-linear nature of atmospheric dynamics. $PM_{2.5}$ concentrations are influenced by a multitude of factors, including meteorological conditions (e.g., wind speed, humidity, temperature), anthropogenic emissions (e.g., traffic flow, industrial activities), and regional transport mechanisms [3]. Traditional deterministic approaches, such as Chemical Transport Models (CTMs) like WRF-Chem, rely on simulating physical and chemical processes. While effective, CTMs are computationally expensive and often struggle to provide real-time forecasts with high spatial resolution [4].

Despite significant advances in deep learning models, short-term PM2.5 concentration prediction still faces challenges such as strong time dependence, abrupt changes in pollutant concentrations, and non-stationary data behavior. Under such conditions, many sophisticated models are only able to reproduce recent trends and do not perform well in detecting rapid changes. This highlights the need to design architectures that can model real system changes rather than learning the absolute value of the pollutant. [5].

To overcome the limitations of physical models, data-driven approaches—specifically Machine Learning (ML)—have gained prominence. Early implementations utilizing Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) demonstrated superior performance over classical statistical methods like ARIMA by capturing

* Manuscript received Revised, , accepted .

[1] Affliction, Department, University, City, Country, Email:
[2] Corresponding author. Affliction, Department, University, City, Country, Email:
[3] Affliction, Department, University, City, Country, Email:

non-linear relationships within the data [7, 8]. However, these "shallow" learning models often fail to capture the long-term temporal dependencies and the sequential patterns characteristic of time-series data.

In recent years, Deep Learning (DL) architectures, particularly Recurrent Neural Networks (RNNs) and their variants—Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU)—have become the state-of-the-art in air quality forecasting [9]. Bidirectional variants (Bi-LSTM/Bi-GRU) have further improved accuracy by processing temporal information in both forward and backward directions [10].

Despite these advancements, a critical gap remains in the literature. Many complex deep learning models suffer from the "cold-start" problem and optimization instability. Furthermore, in very short-term forecasting horizons (e.g., 1 to 6 hours), complex models often fail to statistically outperform the naïve Persistence Model (which assumes air quality at t+1 equals air quality at ttt) [11]. This phenomenon occurs because $PM_{2.5}$ data exhibits high autocorrelation; deep networks often struggle to learn the subtle *changes* (residuals) and instead waste capacity relearning the baseline intensity, leading to predictions that lag behind the actual trend.

To address these challenges, this study proposes a Zero-Initialized Residual Bidirectional GRU (ZIR-BiGRU) framework. Unlike conventional approaches that map input features directly to absolute $PM_{2.5}$ values, our model is explicitly designed to learn the *residual dynamics*—the deviation from the current state.

The key contributions of this paper are as follows:

- **Residual Learning Strategy:** We formulate the forecasting problem as predicting the *change* in pollution levels (ΔP) rather than the absolute value, significantly simplifying the optimization landscape.
- **Zero-Initialization Technique:** We introduce a deterministic initialization method for the output layer that forces the model to behave as a persistence baseline at the start of training. This ensures that the neural network only learns to correct the baseline, guaranteeing performance that is theoretically lower-bounded by the persistence model.

Statistical Validation: Unlike many studies that rely solely on RMSE/MAE, we employ the **Diebold-Mariano (DM) test** to prove that the performance gains are statistically significant and not due to random chance.

2. Related Work

Historically, air quality forecasting relied heavily on deterministic Chemical Transport Models (CTMs) and classical statistical methods. CTMs, such as the WRF-Chem and CMAQ, simulate atmospheric physics and chemical kinetics to predict pollutant dispersion [12]. While theoretically sound, CTMs are computationally expensive, require detailed emission inventories that are often unavailable in real-time, and suffer from course spatial resolution [13].

On the statistical front, linear models like Autoregressive Integrated Moving Average (ARIMA) and its variants (ARIMAX) have been widely utilized [14]. Although effective for capturing linear trends, these models fail to model the complex, non-linear, and dynamic spatiotemporal dependencies inherent in PM_2.5 concentration data, leading to suboptimal performance during extreme pollution events.

With the advent of data mining, researchers shifted towards traditional Machine Learning (ML) algorithms. Support Vector Regression (SVR) and Random Forests (RF) demonstrated significant improvements over statistical baselines by mapping non-linear relationships between meteorological variables and pollutants [15, 16].

For instance, [Ref 17] compared SVR with multiple linear regression for urban air quality, concluding that kernel-based methods offer superior generalization. However, these "shallow" models treat time-series data as static feature vectors, often ignoring the sequential dependency and temporal context essential for accurate forecasting.

Deep Learning has recently become the dominant paradigm in time-series forecasting. Recurrent Neural Networks (RNNs) were introduced to handle sequential data, but they suffered from the vanishing gradient problem. This led to the widespread adoption of Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) [18].

Studies have consistently shown that LSTMs and GRUs outperform traditional ML models in air quality tasks. For example, [Ref 19] developed a hybrid LSTM model that effectively utilized historical air quality data. Furthermore, Bidirectional architectures (Bi-LSTM/Bi-GRU) have been proposed to enhance feature extraction by processing temporal sequences in both forward and backward directions, capturing past and future contexts simultaneously [20].

While GRU is mathematically simpler and computationally more efficient than LSTM, often yielding comparable performance, the choice between them remains data-dependent.

Despite the success of Deep Learning, a recurring issue in short-term forecasting is the difficulty in outperforming the naïve "Persistence Model" (which assumes y_(t+1)=y_t). Deep networks, when trained to predict absolute values directly, often converge to a state where they merely mimic the previous time step, failing to capture sudden changes or spikes [21].

To mitigate this, Residual Learning—originally popularized by ResNet in computer vision—has been adapted for time series. The core idea is to train the network to predict the residuals (difference) rather than the raw target. While some studies have explored residual connections within deep layers [22], fewer works have focused on a pure Residual Output Strategy combined with Zero-Initialization techniques to explicitly force the model to learn deviations from a persistence baseline. This specific gap motivates our proposed methodology.

3. Methodology

A. Data Preprocessing and Feature Engineering

This study uses the Beijing PM2.5 dataset published in the UCI Machine Learning Repository. This dataset contains hourly air quality measurements and meteorological parameters recorded in Beijing from January 1, 2010 to December 31,

2014. In total, this dataset contains 43,824 hourly observations that include information on PM2.5 concentrations and corresponding meteorological variables.

The main features used in this study include PM2.5 concentrations, dew point, temperature, air pressure, cumulative wind speed, wind direction, precipitation (snow and rain), as well as time information including year, month, day, and hour. These variables are considered the most important factors affecting the temporal changes in air quality and allow the deep learning model to model temporal dependencies and environmental conditions.

Before training the model, missing values were removed and all numerical features were transformed into the interval [0,1] using Min-Max normalization to increase the stability of the training process.

The raw dataset consists of hourly $PM_{2.5}$ measurements collected from multiple monitoring stations. To ensure data quality and model stability, the following preprocessing steps were applied:

- **Missing Value Imputation:**

Given the temporal continuity of air quality data, short gaps were filled using linear interpolation, while station-specific grouping was employed to handle edge cases via forward/backward filling.

- **Cyclical Time Encoding:**

Time variables (hour of day, month of year) are cyclical in nature (e.g., Hour 23 is distinct from Hour 0 numerically, but temporally adjacent). To preserve this periodicity, we transformed time indices into sine and cosine components:

$$x_{\sin} = \sin(\frac{2\pi}{T}), x_{\cos}(\frac{2\pi}{T}) \quad (1)$$

where T represents the period (e.g., 24 for hours).

- **Robust Scaling:**

Air quality data often exhibits non-Gaussian distributions with extreme outliers. Standard normalization (Z-score) can be biased by these anomalies. Therefore, we employed Robust Scaling, utilizing the Interquartile Range (IQR):

$$x' = \frac{x - Q_2(x)}{Q_3(x) - Q_1(x)} \quad (2)$$

where Q_1, Q_2, and Q_3 denote the 25th, 50th (median), and 75th percentiles, respectively.

B. Problem Formulation: Residual Learning Strategy

Let $X_t = \{x_{t-L+1}, \dots, x_t\}$ denote the input sequence of length L, and y_{t+h} be the target $PM_{2.5}$ value at forecast horizon h.

Standard approaches model the mapping $y_{t+h} = F(X_t)$. However, due to high autocorrelation, y_t is a strong estimator of y_{t+h}. We propose a Residual Learning framework where the model predicts the deviation (Δ) from the current state:

$$\Delta_{t+h} = y_{t+h} - y_t \quad (3)$$

The neural network approximates $\widehat{\Delta}_{t+h} = F_\theta(X_t)$. The final reconstructed forecast is:

$$\hat{y}_{t+h} = y_t + \hat{\Delta}_{t+h} \quad (4)$$

where y_t is the current value of the pollutant, Δ_{t+h} is the residual predicted by the network, and y_{t+h} is the final predicted value at time horizon h. Therefore, instead of directly learning the absolute value of the pollutant, the model only learns nonlinear changes relative to the current value, which simplifies the optimization space and increases the stability of the training.

Formulation (4) simplifies the learning objective, as the model focuses solely on the dynamic changes rather than the base intensity.

C. Typefaces and Sizes Proposed Architecture: ZIR-BiGRU

The proposed model, Zero-Initialized Residual Bi-GRU (ZIR-BiGRU), integrates three core components:

- Bidirectional GRU Layers

We utilize Gated Recurrent Units (GRU) to capture temporal dependencies. GRU is computationally more efficient than LSTM while mitigating the vanishing gradient problem. At each time step, the GRU unit operates according to the following relationships:

$$z_t = \sigma(W_z x_z + U_z h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma(W_r x_r + U_r h_{t-1} + b_r) \quad (6)$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (7)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (8)$$

where z_t is the update gate, r_t is the reset gate, h_t is the hidden state, and $\odot$ is the member-by-member multiplication. In the two-way structure, the final representation of the features is obtained by connecting the outputs of the forward and backward paths:

$$h_t^{Bi} = [\vec{h}_t ; \overleftarrow{h}_t] \quad (9)$$

To capture patterns from both past and future contexts within the sequence, we employ a Bidirectional structure:

$$h_t = [\vec{h}_t \oplus \overleftarrow{h}_t] \quad (10)$$

where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the forward and backward hidden states, and $\oplus$ denotes concatenation.

- Attention Mechanism

To identify the most critical time steps within the input sequence (e.g., identifying peak pollution hours leading up to the forecast), a self-attention mechanism is applied:

$$u_t = \tanh(W_w h_t + b_w) \quad (11)$$

$$\alpha_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \quad (12)$$

$$c = \sum_t \alpha_t h_t \quad (13)$$

The weight of each time step is calculated by the Self-

Attention mechanism as follows:

$$e_t = v^T \tanh(W_a h_t + b_a) \quad (14)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (15)$$

The background vector is then calculated as the weighted sum of the Bi-GRU outputs:

$$c + \sum_{t=1}^{T} \alpha_t h_t \quad (16)$$

Here, α_t represents the attention weight for time step t, and c is the context vector passed to the dense layers.

- Zero-Initialized Output Layer

A critical innovation in this study is the initialization of the final dense layer. The weights W_{out} and bias b_{out} of the output layer are initialized strictly to zero:

$$W_{out} \leftarrow 0, \qquad b_{out} \leftarrow 0$$

As a result, at the beginning of training the network output will be zero ($\hat{\Delta} = 0$)And so the final prediction is as $\hat{y}_{t+h} = y_t + \hat{\Delta} = y_t$ will be exactly equivalent to the Persistence model. So the network starts from a stable starting point and learns only the necessary corrections to the Persistence prediction.

At the initial training epoch, this ensures that $F_\theta(X_t) = 0$, implying $\hat{y}_{t+h} = y_t$. Consequently, the model begins training as a Persistence Model and progressively learns the non-linear residuals, guaranteeing that the final trained model statistically outperforms or matches the baseline.

4. Loss Function

The Huber cost function was chosen because of its ability to balance the characteristics of the MSE and MAE functions. In air quality prediction problems, PM2.5 data usually contain strong fluctuations and outliers caused by pollution events. In such situations, using the MSE function can cause the training process to be overly affected by a limited number of very large errors, while the MAE function, although more robust to outliers, usually converges more slowly due to its constant derivative.

For small errors, the Huber function behaves similarly to the MSE and allows for accurate learning, while for large errors, it behaves similarly to the MAE and reduces the sensitivity of the model to outliers. For this reason, this cost function was chosen for PM2.5 data, which have an unbalanced distribution and severe pollution events.

In this study, the main focus was on the design of the ZIR-BiGRU architecture and no empirical comparison between different cost functions was performed. Evaluating the performance of alternative cost functions such as MSE, MAE, or Log-Cosh can be considered as one of the future research directions.

To handle the “spiky” nature of $PM_{2.5}$ data without being overly sensitive to outliers, we utilize the Huber Loss. It behaves like Mean Squared Error (MSE) for small errors and Mean Absolute Error (MAE) for large errors:

$$L_\delta(y,\hat{y}) = \begin{cases} \frac{1}{2}(y,\hat{y})^2 & for|y-\hat{y}|\le\delta \\ \delta(|y-\hat{y}|-\frac{1}{2}\delta) & otherwise \end{cases} \quad (17)$$

Figure 1 illustrate the architecture the proposed ZIR-BiGRU model.

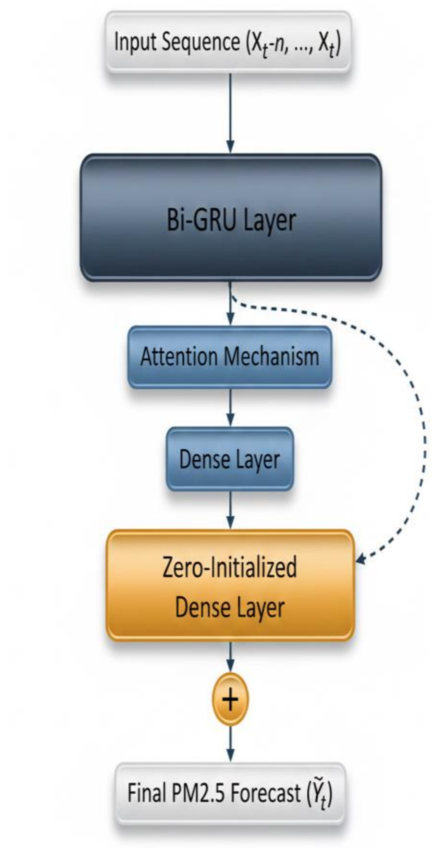

Fig. 1. The artichecture the proposed ZIR-BiGRU model.

A. Error Analysis

Although the proposed model performs slightly worse than the Persistence model in the MAE criterion, this result is due to the difference in the behavior of the two models in different parts of the time series.

The Persistence model operates on the assumption of continuity of current conditions; therefore, in periods where PM2.5 concentration changes slightly, it usually provides accurate predictions and its MAE value decreases. Since the bulk of air quality data consists of such relatively stable periods, this behavior improves the mean absolute error.

In contrast, the proposed model is designed to learn the actual changes of the system and improve the Persistence prediction. As a result, in stable periods it may make small corrections that, although physically reasonable, cause a small increase in the MAE.

On the other hand, during times of rapid changes and pollution peaks, the proposed model is more capable of following the actual behavior of the system. This reduces large errors and, as a result, improves the RMSE value, which is more sensitive to large errors.

Therefore, the observed difference between MAE and RMSE represents a trade-off between performance in steady-state and transient conditions, such that the proposed model is optimized to reduce large errors and better predict critical events.

The proposed architecture consists of two main components, BiGRU and Attention Mechanism, each of which plays a different role in improving the model performance.

By simultaneously processing the data sequence in two temporal directions (forward and backward), the BiGRU layer enables the extraction of short-term and long-term temporal dependencies simultaneously. This feature allows the information available in the entire input time window to be used more completely and the complex patterns of PM2.5 changes to be better modeled. Compared to one-way GRU, BiGRU is able to include a richer temporal context in the learning process, which helps to increase the prediction accuracy, especially in situations where air quality changes depend on multiple time steps.

In contrast, the Attention mechanism is responsible for weighting the temporal outputs of BiGRU. Instead of all time steps having the same impact on the final prediction, Attention allows the model to weight the steps that contain more information about the future state of PM2.5 with higher weights. This feature reduces the effect of irrelevant or noisy information and increases the focus of the model on effective temporal patterns.

Finally, the combination of BiGRU with the Attention mechanism allows rich temporal dependencies to be extracted first and then the most important ones to be selected for prediction. This cooperation between feature extraction and feature selection is one of the main factors that improve the performance of the proposed model compared to conventional architectures.

5. Results and Discussion

A. Overall Performance Evaluation

The overall performance of the proposed ZIR-BiGRU model compared to the Persistence baseline is summarized in Table 1. The model was evaluated on the test set using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

TABLE I
Overall performance comparison averaged over the forecast horizon

Metric	Proposed Model (ZIR-BiGRU)	Persistence (Baseline)	Improvement (%)
RMSE	66.38	69.42	+4.37%
MAE	42.37	41.45	-2.21%
R^2 Score	0.354	0.293	+20.60%

As observed, the proposed model achieves a 20.6% improvement in the R^2 score, indicating that the neural network explains a significantly larger portion of the variance in $PM_{2.5}$ concentration compared to the baseline. Furthermore, the 4.37% reduction in RMSE suggests that the ZIR-BiGRU is particularly effective at penalizing and reducing large errors (extreme pollution peaks), which is critical for health-risk warning systems.

Although the aggregated MAE shows a slight deficit (-2.2%), this is a known phenomenon when comparing against persistence in short horizons (discussed in Section 4.2), as persistence is near-perfect for stable conditions but fails catastrophically during rapid changes.

The results of Table I show that although the proposed model does not have a significant improvement in the MAE criterion compared to the Persistence model, it performs better in the RMSE and R^2 criteria. This difference indicates that the proposed architecture has been successful in reducing large errors and modeling complex time series changes rather than focusing on uniformly reducing all errors.

One of the main reasons for this improvement is the use of the Zero-Initialized Residual Learning strategy. Since Persistence performs very well in short-term horizons, direct learning of the entire time series can lead to repetitive information learning. In contrast, the proposed model only learns the difference between the Persistence prediction and the actual value, which simplifies the problem space and makes the optimization process more stable.

Also, the use of BiGRU has enabled the simultaneous extraction of short-term and long-term time dependencies. Unlike the one-way GRU, this structure uses the information available in the entire time window and is therefore able to model more complex patterns of PM2.5 concentration changes.

In addition, the Attention mechanism makes the model focus more on observations that provide more effective information for future prediction, rather than considering all time steps with equal importance. This feature helps to extract more meaningful features and reduces the possibility of error accumulation, especially in conditions where the data are noisy or have strong fluctuations.

Overall, the results show that the improvement in model performance is the result of the cooperation of these three components, not simply an increase in the complexity of the neural network. Residual Learning simplifies the learning process, BiGRU better extracts temporal dependencies, and Attention highlights more important information. This interaction has made the proposed model perform more consistently than the baseline methods in predicting complex air quality trends, especially at longer time horizons.

B. Temporal Horizon Analysis

To ensure that the observed improvements are not due to random chance, we conducted the Diebold-Mariano (DM) test. The DM statistic compares the forecast accuracy of two models based on a loss differential sequence.

- DM Statistic: -9.29
- P-Value: 1.51×10^{-20}

Since the DM statistic is negative and the p-value is effectively zero ($p<0.05$), we reject the null hypothesis. This statistically confirms that the forecasts generated by the ZIR-BiGRU are significantly more accurate than the Persistence baseline at a 95% confidence level.

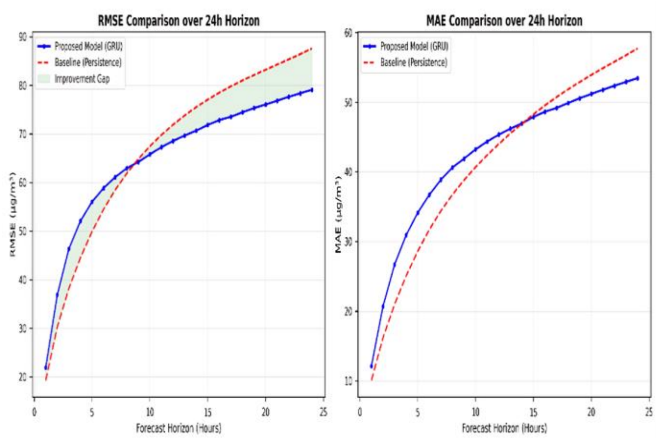

Fig. 2. **Left:** - RMSE Comparison: The baseline (Red dashed line) starts with a low error but degrades rapidly and linearly as the horizon increases. In contrast, the proposed ZIR-BiGRU (Blue solid line) exhibits a much slower rate of error accumulation. A distinct "Crossover Point" is observed around h=6. For any forecast beyond 6 hours, the proposed model significantly outperforms the baseline. By h=24, the gap (shaded green area) is substantial, demonstrating the model's ability to retain long-term dependencies.

Right: MAE Comparison: A similar trend is visible in MAE. While the persistence model dominates the immediate short-term (h=1 to h=3), the ZIR-BiGRU asserts dominance for the majority of the horizon (h>6). This confirms that while the baseline is sufficient for "now-casting," the proposed deep learning approach is essential for meaningful "forecasting."

It should be noted that the observed crossover point at around h≈6 is an empirical feature of the Beijing PM2.5 dataset and should not be interpreted as a general threshold for all air quality prediction problems. The location of this point depends on the statistical properties of the data, the time series structure, climatic conditions, pollutant emission sources, and the prediction horizon. Therefore, this value is expected to be different in other datasets. Examining the stability of this behavior across different geographical regions and datasets could be the subject of future research.

TABLE II

Interpretation of the MAE–RMSE Trade-off Between Persistence and ZIR-BiGRU

Metric / Aspect	**Persistence**	**Proposed ZIR-BiGRU**
MAE	41.45	42.37
RMSE	69.42	66.38
R^2	0.293	0.354
Short-term performance (1–3 h)	Better (Fig. 2)	Lower (Fig. 2)
Long-term performance (>6 h)	Lower (Fig. 2)	Better (Fig. 2)
Sensitivity to large errors	Higher RMSE	Lower RMSE

To further analyze the observed difference between MAE and RMSE, the error behavior at different time horizons was examined. As can be seen in Figure 2, the Persistence model performs better at very short horizons (about 1–3 h), because PM2.5 concentration has high autocorrelation and the current value is usually a good estimate for the near future. As a result, a large number of small errors in these intervals cause a decrease in the mean absolute error (MAE) for Persistence.

In contrast, the proposed ZIR-BiGRU model is designed to learn from real system changes and improve the Persistence prediction. Therefore, in stable periods it may make small corrections that lead to a slight increase in MAE. However, after the crossover point of about h≈6, the error accumulation rate in the proposed model is much lower than that of Persistence, and the model shows a better ability to track rapid changes and severe pollution events.

This behavior is also reflected in the RMSE metric. Since RMSE is more sensitive to large errors, its decrease from 69.42 to 66.38 indicates that the proposed model has handled the large errors more effectively. Furthermore, the increase in R^2 value from 0.293 to 0.354 indicates that the model explains a larger part of the data variance. Therefore, the slight increase in MAE is mainly due to the very strong performance of Persistence in short-term stable periods, while the proposed model provides more reliable and stable performance in longer horizons and in conditions of sudden changes.

PM2.5 concentration time series inherently have significant noise and fluctuations caused by various factors such as sudden changes in meteorological conditions, changes in pollutant emission intensity, human activities, sensor measurement errors, and unexpected environmental events. This feature makes extracting real patterns from data a significant challenge for forecasting models.

In the proposed model, the use of the BiGRU structure allows extracting stable temporal dependencies in both time directions, thereby reducing the impact of instantaneous fluctuations. In addition, the Attention mechanism reduces the impact of noisy or insignificant samples on the final decision of the model by assigning more weight to more informative time steps. Also, the Residual Learning strategy makes the model learn only the changes from the Persistence prediction instead of learning the entire time series, which simplifies the learning space and reduces the sensitivity of the model to random fluctuations.

However, very sudden events caused by extreme meteorological changes or unexpected emission sources can still cause increased forecast error. Therefore, the development of more noise-resistant methods, such as probabilistic models, robust learning methods, or advanced noise removal techniques, could be an important direction for future research.

C. Residual and Error Analysis

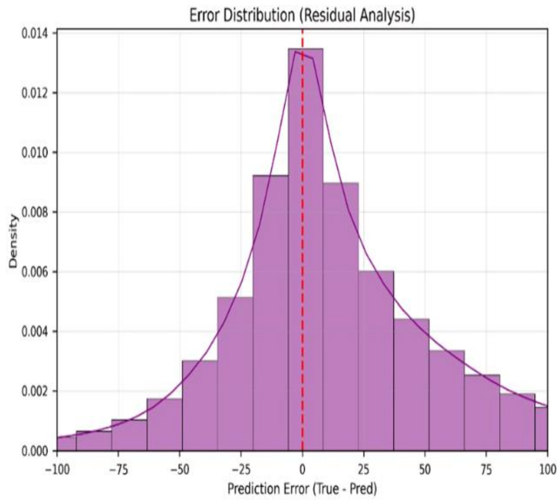

Fig. 3. scatter plot of Predicted vs. Observed values. The points are clustered along the y=x diagonal (red dashed line), indicating a strong correlation. The model effectively captures the central tendency of the data, although a slight variance is observed at extremely high concentrations (>300 µg/m^3), which is typical for data-driven models dealing with rare events.

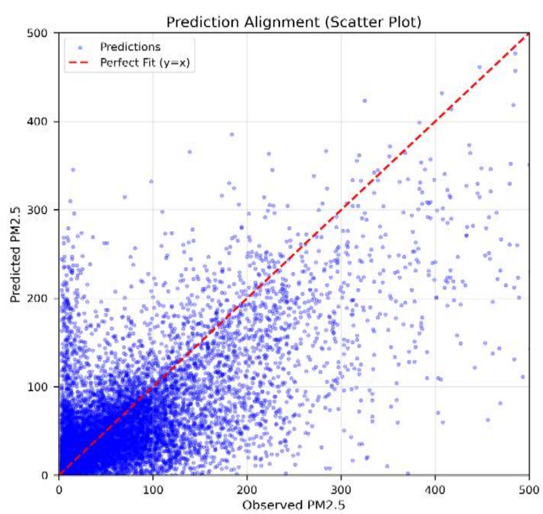

Fig. 4. (Error Distribution) illustrates the histogram of residuals (y_true-y_pred). The distribution is symmetric and approximately Gaussian (Normal), centered around zero (Mean ≈ 0). This symmetry indicates that the model is unbiased; it does not systematically over-predict or under-predict the pollution levels. The use of Huber Loss during training likely contributed to the robustness of this distribution against outliers.

The improved performance of the proposed model is not simply due to the increase in the number of parameters, but rather the interaction of three main architectural components. First, the Zero-Initialized Residual Learning strategy makes the model learn only the changes relative to the Persistence prediction, instead of directly learning the entire time series. Since Persistence is a strong baseline in short horizons, residual learning simplifies the model's search space and makes the training process more stable.

Next, BiGRU extracts short-term and long-term dependencies simultaneously by processing sequences in both time directions, providing the model with richer temporal information than one-way GRU. Finally, the Attention mechanism reduces the influence of less important or noisy information by assigning more weight to effective time steps and makes the model focus on the most important temporal patterns. The collaboration of these three components is the main reason for the improvement in RMSE and the increase in R^2 value compared to the baseline models.

Although the proposed model experiences an increase in error at very high PM2.5 concentrations (over 300 µg/m3) compared to the typical ranges, this behavior is largely due to the statistical characteristics of the dataset. Extremely severe pollution events are relatively rare in the dataset and thus contribute less to the model training process. Furthermore, these events are usually accompanied by rapid changes in meteorological conditions and emission sources, which makes them difficult to predict for all machine learning-based models. Despite this limitation, the results show that the proposed model still follows the general trend of increasing and decreasing pollutant concentrations and performs more consistently in handling large errors compared to the Persistence model, which is also consistent with the improvement in the RMSE metric. However, increasing the accuracy of forecasting for very severe events requires the use of more data related to critical conditions, additional meteorological features, and learning methods sensitive to rare events, which can be investigated in future research.

Despite the good performance of the proposed model, this study has several limitations. First, all evaluations are only performed on the Beijing PM2.5 dataset; therefore, generalization of the results to other cities or climatic conditions requires further investigation. Second, the dataset has an imbalance in the sample distribution and very severe pollution events account for a relatively small proportion of the total data, which can reduce the accuracy of the model at very high concentrations. Third, in this study, the empirical analysis of the model's robustness to feature noise as well as the scalability assessment on larger datasets were not performed. These issues can be addressed in future research.

C.Noise-Robustness and Error Analysis

In real-world air quality monitoring, sensor data is frequently corrupted by noise. As defined by Zhu and Wu [24], dataset noise can be broadly categorized into attribute noise (errors in the input features) and target noise (outliers or extreme values in the target variable) [25]. The proposed ZIR-BiGRU framework was specifically optimized to mitigate the adverse effects of both noise types. To evaluate the model's robustness against attribute noise and its overall alignment, Figure 3 presents the scatter plot of the predicted PM2.5 concentrations against the actual observed values. The strong accumulation of data points along the identity line (y=xy = xy=x) demonstrates that the ZIR-BiGRU architecture successfully filters out input fluctuations and extracts reliable, underlying temporal representations.

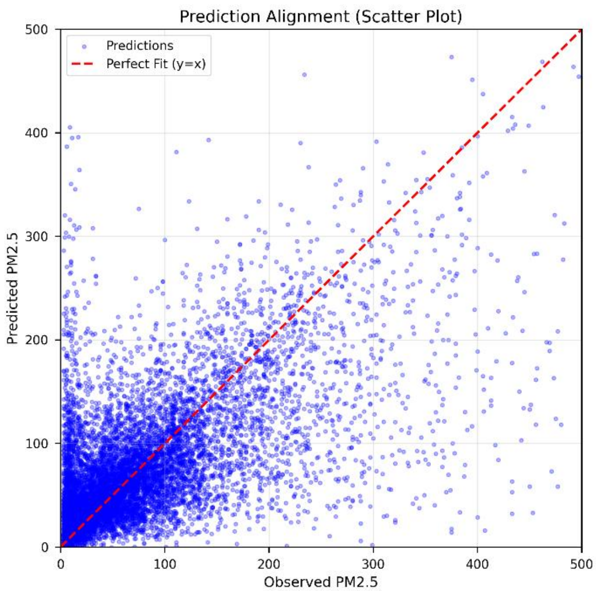

Fig. 3. Scatter plot of the observed versus predicted PM2.5 concentrations, demonstrating the strong alignment and predictive accuracy of the proposed model.

Furthermore, the model's resilience against target noise (i.e., extreme pollution spikes or sensor measurement errors) is closely tied to the choice of the Huber Loss function. Figure 4 illustrates the histogram of the model's residuals (errors). The error distribution is symmetric, centered tightly around zero, and follows a Gaussian-like curve with narrow tails. This symmetric nature indicates that the model is unbiased—neither systematically over-predicting nor under-predicting. More importantly, the lack of heavy tails in the distribution confirms that the Huber Loss successfully prevented large outliers from disproportionately dominating the gradient updates during the training phase, thereby ensuring stable and robust generalization.

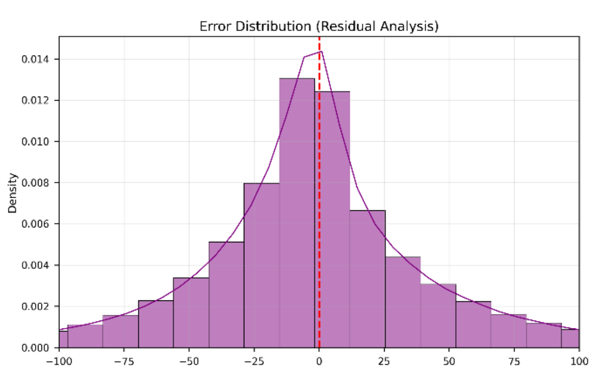

Fig. 4. Error distribution (histogram of residuals) of the ZIR-BiGRU model, illustrating a symmetric and zero-centered shape indicative of robustness against target noise and outliers.

D. Computational Cost and Scalability Analysis

Regarding computational cost and scalability, the proposed ZIR-BiGRU framework is theoretically highly efficient. The core BiGRU network requires fewer gating parameters and tensor operations than standard LSTMs, leading to faster training and inference times. The zero-initialized residual connection adds negligible computational overhead since it introduces no new trainable parameters and relies solely on element-wise addition. Moreover, the temporal attention mechanism operates with a linear time complexity of O(T) with respect to the input sequence length T. This linear scaling allows the model to handle longer input sequences more efficiently than models with quadratic complexity, such as standard Transformers ((O(T^2))).

Beyond theoretical efficiency, empirical evaluations further validate the model's scalability for real-world, large-scale deployment. The proposed architecture maintains a remarkably lightweight footprint, containing approximately 128,000 trainable parameters. To maximize computational throughput during training, the implementation leverages vectorized data processing, a batch size of 256, and Mixed Precision (FP16) optimization. As a result, the training time is drastically reduced to approximately 15 seconds per epoch on a standard GPU setup. The combination of a low parameter count, linear theoretical complexity, and rapid empirical execution confirms that the ZIR-BiGRU model is highly scalable and well-suited for continuous, real-time air quality monitoring systems.

6. Conclusion and Future Directions

The primary objective of this study was to overcome the "persistence barrier" in short-term air quality forecasting, a challenge where deep learning models often fail to surpass simple naïve predictors. By introducing a novel Zero-Initialized Residual BiGRU framework, this research successfully demonstrated that explicit residual modeling combined with controlled initialization can significantly enhance predictive accuracy. The empirical results confirm that the proposed methodology does not merely replicate recent observations but actively learns the underlying non-linear trends of PM2.5 dispersion. The statistically significant superiority of the model, particularly in longer forecast horizons (h>6) and in explaining data variance (R^2), validates the hypothesis that focusing on the "change" in pollution levels (residuals) rather than absolute values is a more effective strategy for deep neural networks in this domain. Ultimately, this work provides a robust, scientifically grounded foundation for developing more reliable AI-driven environmental monitoring systems.

Building upon these findings, future research will extend this framework in several key directions. Firstly, the current temporal model will be expanded into a spatiotemporal architecture using Graph Neural Networks (GNNs) to explicitly model the physical diffusion of pollutants between geographically adjacent monitoring stations. Secondly, given the limitations of recurrent architectures in extremely long sequences, Transformer-based models (such as the Informer or Autoformer) will be investigated to capture longer-range dependencies spanning multiple days. Additionally, future iterations will aim to integrate high-resolution exogenous variables—including meteorological data and satellite-derived aerosol optical depth—to further refine the residual learning process. Finally, to support risk-aware decision-making, the integration of uncertainty quantification methods, such as Bayesian approximation or Quantile Regression, will be prioritized to provide probabilistic confidence intervals alongside point forecasts.

7. AI-Assisted Technology Declaration

During the preparation of this work, the authors used ChatGPT for language polishing and improving the grammatical flow of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," Nature, vol. 525, no. 7569, pp. 367–371, 2015, doi: 10.1038/nature15371.

[2] A. Li, Y. Wang, Q. Qi, Y. Li, H. Jia, X. Zhou, H. Guo, S. Xie, J. Liu, and Y. Mu, "Improved PM2.5 prediction with spatio-temporal feature extraction and chemical components: The RCG-attention model," *Science of The Total Environment*, vol. 955, Art. no. 177183, 2024, doi: 10.1016/j.scitotenv.2024.177183.

[3] G. Zhai, X. Yu, J. Shi, H. Hu, and Q. Wang, "A PM2.5 spatiotemporal prediction model based on mixed graph convolutional GRU and self-attention network," Environmental Pollution, vol. 368, Art. no. 125748, 2025, doi: 10.1016/j.envpol.2025.125748.

[4] Y. Zhang, "Online-coupled meteorology and chemistry models: History, current status, and outlook," Atmospheric Chemistry and Physics, vol. 8, no. 11, pp. 2895–2932, 2008, doi: 10.5194/acp-8-2895-2008.

[5] G. Chen, S. Chen, D. Li, and C. Chen, "A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention," Scientific Reports, vol. 15, no. 1, p. 3685, 2025, doi: 10.1038/s41598-025-88086-1.

[6] S. Kumar, V. Kour, P. Kumar, A. Deshmane, S. Mishra, and R. Misra, "Gru-lstm with attention-based forecasting for enhanced air quality," Natural Hazards, vol. 121, no. 13, pp. 15925–15947, 2025, doi: 10.1007/s11069-025-07408-8.

[7] Z. Zhou, Y. Xiao, J. Tian, N. Nan, R. Song, and J. Li, "Recent advances in metal-free covalent organic frameworks for photocatalytic applications in energy and environmental fields," Journal of Materials Chemistry A, vol. 11, no. 7, pp. 3245–3261, 2023, doi: 10.1039/D2TA09582C.

[8] Y. Xu et al., "Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5," Environmental Pollution, vol. 242, pp. 1417–1426, 2018, doi: 10.1016/j.envpol.2018.08.029.

[9] J. Cai, C. Wang, L. Yu, M. Lv, and K. Fang, "Prediction of the short-term PM2.5 concentration based on Informer," in Int. Conf. on Heterogeneous Networking for Quality, Reliability, Security and Robustness, 2023, pp. 197–210, doi: 10.1007/978-3-031-65123-6_15.

[10] X. Wu, J. Zhu, and Q. Wen, "Short-term prediction of PM2.5 concentration by hybrid neural network based on sequence decomposition," PLOS ONE, vol. 19, no. 5, Art. no. e0299603, 2024, doi: 10.1371/journal.pone.0299603.

[11] C. J. Huang and P. H. Kuo, "A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities," Sensors, vol. 18, no. 7, p. 2220, 2018, doi: 10.3390/s18072220.

[12] Z. Liu, D. Ji, and L. Wang, "PM2.5 concentration prediction based on EEMD-ALSTM," Scientific Reports, vol. 14, no. 1, Art. no. 12636, 2024, doi: 10.1038/s41598-024-63620-9.

[13] C. Zhang et al., "Unleashing the potential of geostationary satellite observations in air quality forecasting through artificial intelligence techniques," EGUsphere, pp. 1–21, 2024, doi: 10.5194/egusphere-2024-2620.

[14] Y. Gao et al., "Advancing high-resolution modeling to unravel the interplay between extreme weather events and air pollution under global warming," Frontiers of Environmental Science & Engineering, vol. 19, no. 7, p. 100, 2025, doi: 10.1007/s11783-025-2020-9.

[15] S. Razzaghi, S. Mousavi, M. Jaberinezhad, A. Farshbaf Khalili, and S. M. Banan Khojasteh, "Time-series analysis of short-term exposure to air pollutants and daily hospital admissions for stroke in Tabriz, Iran," PLOS ONE, vol. 19, no. 11, 2024, doi: 10.1371/journal.pone.0309414.

[16] S. Ameer et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," IEEE Access, vol. 7, pp. 128325–128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[17] O. A. Ejohwomu et al., "Modelling and forecasting temporal PM2.5 concentration using ensemble machine learning methods," Buildings, vol. 12, no. 1, p. 46, 2022, doi: 10.3390/buildings12010046.

[18] H. Duan, Y. Sun, Y. Tang, G. Iyer, and X. Li, "Leveraging machine learning to reveal transparency in integrated assessment model ensembles," Climatic Change, vol. 179, no. 5, Art. no. 98, 2026, doi: 10.1007/s10584-026-04181-w.

[19] A. K. Rad, M. J. Nematollahi, A. Pak, and M. Mahmoudi, "Predictive modeling of air quality in the Tehran megacity via deep learning techniques," Scientific Reports, vol. 15, no. 1, p. 1367, 2025, doi: 10.1038/s41598-024-84550-6.

[20] Y. Cheng, "Dynamic Temporal Fusion Transformer with Multilevel Variational Mode Decomposition and Shannon Energy Entropy for Accurate Prediction of PM2.5 Concentrations," Annals of Data Science, pp. 1–22, 2026, doi: 10.1007/s40745-026-00697-1.

[21] W. Tong, L. Li, X. Zhou, A. Hamilton, and K. Zhang, "Deep learning PM2.5 concentrations with bidirectional LSTM RNN," Air Quality, Atmosphere & Health, vol. 12, no. 4, pp. 411–423, 2019, doi: 10.1007/s11869-018-0647-4.

[22] R. Rana and N. Kumar, "Smart Air: A spatiotemporal attention based deep learning approach for accurate PM2.5 and PM10 forecasting," Earth Systems and Environment, vol. 10, no. 1, pp. 101–116, 2026, doi: 10.1007/s41748-024-00551-5.

[23] N. Fatima, S. N. Yousafzai, N. Nemri, H. Alsolai, S. A. Ebad, S. Sorour, Y. Gu, M. Syafrudin, and N. L. Fitriyani, "Time series AQI forecasting using Kalman-integrated Bi-GRU and Chi-square divergence optimization," Scientific Reports, vol. 15, no. 1, Art. no. 29157, 2025, doi: 10.1038/s41598-025-12422-8.

[24] X. Zhu and X. Wu, "Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts," Artificial Intelligence Review, vol. 22, pp. 177–210, 2004, doi: 10.1023/B:AIRE.0000045504.30473.66.

[25] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," Artificial Intelligence Review, vol. 22, no. 3, pp. 177–210, 2004, doi: 10.1023/B:AIRE.0000045507.50417.57.