

Article

Flipout Bayesian LSTM with Residual Attention for Uncertainty-Aware PM_{2.5} Forecasting and Anomaly Detection

Quan Li ¹, Huaxing Lu ^{1,*}, Haiyang Xu ² and Dengwei Sun ¹¹ School of Geomatics Science and Technology, Nanjing Tech University, Nanjing 211816, China; nanxi@njtech.edu.cn (Q.L.); sdw@njtech.edu.cn (D.S.)² School of Statistics and Data Science, LEBPS, KLMDASR, and LPMC, Nankai University, Tianjin 300071, China; xuhaiyang@mail.nankai.edu.cn

* Correspondence: luhuaxing@njtech.edu.cn

Abstract

Accurate PM_{2.5} prediction and reliable uncertainty assessments are essential for effective early warnings and public health protection. However, most existing deep learning models only provide deterministic predictions, with limited treatment of predictive uncertainty, which may reduce the reliability under noisy or abrupt pollution conditions. This study presents a flipout Bayesian LSTM with residual attention (FBA-LSTM), which integrates Bayesian flipout inference, residual connections, and temporal attention to jointly improve the predictive accuracy and uncertainty estimation. Unlike MC dropout, our model explicitly represents weight distributions through variational flipout inference, yielding more comprehensive and stable uncertainty estimates with a lower computational cost. A lightweight calibration module based on standard-deviation scaling further aligns the confidence intervals with empirical coverage. Experiments on hourly PM_{2.5} data from four Nanjing stations (2020) showed that FBA-LSTM improves the accuracy, noise robustness, and exceedance warnings ($F_1 = 0.996$) and achieves a higher anomaly-detection performance ($F_1 = 0.691$) than baseline models and methods, thereby facilitating the realization of urban environmental sustainability and public health security.

Keywords: PM_{2.5} forecasting; uncertainty quantification; Bayesian inference; model calibration; anomaly detection

1. Introduction

Air pollution has emerged as a critical global health issue. Fine particulate matter (PM_{2.5}), a major air pollutant, is strongly associated with numerous respiratory and cardiovascular diseases [1–3]. Accurately predicting PM_{2.5} concentrations and quantifying their uncertainty can provide timely and reliable health warnings to the public, prevent false alarms and missed warnings, and avoid the waste of public resources. This integrated prediction, quantification, and warning system is essential for evidence-based environmental governance and guiding the public to take targeted travel precautions [4].

Currently at the city scale, early warning systems are commonly organized through forecast-triggered warnings and response actions, which places high demands on the short-term PM_{2.5} prediction performance. For example, Lv et al. developed a deterministic non-linear empirical regression model for next-day PM_{2.5} forecasting in Nanjing and deployed it within an automated data-retrieval and operational forecasting platform [5], demonstrating that short-term forecasting models can be embedded into routine early warning service



Academic Editor: Giouli Mihalakakou

Received: 13 January 2026

Revised: 2 February 2026

Accepted: 5 February 2026

Published: 7 February 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

systems. Beyond ground-based forecasting, Earth-observation-based remote sensing has become critical for air quality early warnings, especially in large-scale, sparsely monitored regions. For instance, Semlali et al. proposed a near-real-time monitoring architecture integrating heterogeneous satellite data (e.g., from NASA/NOAA/ESA) with complex event processing (CEP): it preprocesses multi-source observations via SAT-ETL-Integrator, computes AQI in seconds using SAT-CEP-Monitor, and generates interactive AQI maps for visualization. Validated against Spanish ground-station data, this remote-sensing system serves as a rapid, wide-coverage supplement to ground-based early warning networks [6,7]. Nevertheless, achieving accurate PM_{2.5} predictions with reliable uncertainty quantification and early warning capabilities remains challenging.

The existing methods for PM_{2.5} concentration prediction, encompassing statistical models, machine learning techniques, and deep learning models, all have limitations in uncertainty quantification. Statistical models, including the autoregressive integrated moving average model (ARIMA) [8], multiple linear regression (MLR) [9], and the Kalman filter, excel at capturing the linear temporal patterns of PM_{2.5}, but struggle to model complex nonlinear relationships, such as the intricate interactions between the PM_{2.5} concentration, meteorology, and socioeconomic development. Machine learning models, like random forest [10] and support vector machines (SVM) [11], address nonlinearity to some extent, yet face high computational costs and overfitting issues when handling large-scale hourly air quality data. Deep learning models, particularly long short-term memory networks (LSTMs) and their attention-enhanced variants [12–14], have emerged as the dominant approach in recent years due to their ability to capture long-term temporal dependencies. Nevertheless, a common flaw persists among these methods: they only offer deterministic point predictions and lack a systematic framework for quantifying the prediction uncertainty, severely limiting their reliability in real-world scenarios with unexpected pollution events or noisy data [15].

To address this gap, two types of uncertainty-aware frameworks have been proposed for LSTM-based PM prediction, but each has critical drawbacks. Approximate Bayesian methods, such as Monte Carlo (MC) dropout LSTM [16], simulate predictive uncertainty by randomly deactivating neurons during forward passes and aggregating the resulting stochastic outputs. While computationally feasible, MC dropout relies on local perturbations within a fixed network structure, failing to capture global parameter posterior distributions or interdependencies, thereby leading to imprecise and non-systematic uncertainty estimates [17]. Traditional Bayesian inference approaches, such as Bayesian LSTM [18], explicitly model parameter uncertainty through posterior estimation, providing a more comprehensive quantification of predictive uncertainty. However, they face prohibitive computational costs: Variational inference (VI) struggles with prior selection and optimization in complex architectures [19], while Markov chain Monte Carlo (MCMC) methods require massive samples and are impractical for large-scale hourly datasets [20]. The flipout method [21] has emerged as a more efficient alternative to traditional Bayesian approaches. It uses pseudo-independent sampling with shared base perturbations to decouple sample gradients, reducing the computational redundancy. By performing mini-batch computations via matrix multiplication, flipout fully leverages GPU parallelism, accelerating the training speed and improving efficiency, especially for large-scale datasets [22].

Despite these advancements, three critical gaps remain in current PM_{2.5} predictions with uncertainty quantification: (1) MC dropout and similar approximate methods only provide local, non-systematic uncertainty estimates that do not reflect global parameter uncertainties, leading to unreliable confidence in extreme pollution events; (2) traditional Bayesian methods (VI/MCMC) offer rigorous uncertainty quantification, but are computationally prohibitive for large-scale temporal air quality data, limiting real-time application;

and (3) existing uncertainty-aware models rarely integrate feature enhancement mechanisms capable of capturing the key temporal dependencies of PM_{2.5} concentrations, leading to a suboptimal predictive accuracy and unreliable uncertainty estimates. These gaps motivate the need for a computationally efficient, systematically rigorous, and accuracy-enhanced framework for PM_{2.5} predictions with uncertainty quantification.

Building upon this foundation, we propose a PM_{2.5} prediction model, flipout Bayesian LSTM with residual attention (FBA-LSTM), that integrates a Bayesian LSTM with residual attention mechanisms to jointly enhance noise resistance, anomaly detection, and the prediction accuracy. Through Bayesian inference implemented via the flipout method, the model produces not only predictive values, but also their associated confidence intervals, thereby improving the reliability and decision-support capability. Moreover, the combination of attention and residual mechanisms enables the model to effectively capture salient temporal–spatial features while mitigating the performance degradation caused by the network depth, resulting in a higher training efficiency and feature utilization. Finally, a post-processing calibration approach [23] is employed to refine the predicted probability distributions, further improving calibration and ensuring a close alignment between the predicted and observed values. Comparative experiments with classical baselines—including LSTM, attention LSTM, residual LSTM and MC dropout LSTM—demonstrated the robustness and superior predictive performance of the proposed model under diverse pollution scenarios. Our main contributions are as follows:

- (1) We propose a PM_{2.5} prediction model (FBA-LSTM) that integrates flipout-based Bayesian inference, LSTM, and residual attention mechanisms to enhance the accuracy, noise resistance, and anomaly detection. Flipout reduces the computational cost of Bayesian inference, while residual attention highlights noise-obscured key features.
- (2) We employed the flipout method for Bayesian inference to tackle traditional Bayesian methods' computational and training challenges. Flipout reduces gradient variance through shared perturbations, enabling efficient uncertainty quantification for large-scale hourly PM_{2.5} datasets.
- (3) We incorporated a post-processing calibration procedure to refine the output probability distributions and improve the prediction confidence. This step mitigates the overconfidence bias in MC dropout-based models, aligning the predictive intervals with observed errors and improving the reliability of uncertainty estimates in pollution forecasting.

The remainder of this paper is organized as follows. Section 2 describes the study area, data sources, data preprocessing procedures, and overall methodological framework, including the proposed flipout Bayesian LSTM with residual attention, model calibration strategies, and training settings. Section 3 presents the experimental results, including comparisons of different calibration methods, the prediction accuracy, the early warning performance, and the outlier detection capability. Section 4 discusses the results in the context of previous studies, analyzes the computational cost and convergence characteristics, and highlights practical challenges and future research directions. Finally, Section 5 concludes the paper and summarizes the main findings and contributions.

2. Materials and Methods

2.1. Study Area

This study selected Nanjing (32°03' N, 118°46' E) as the study area. Nanjing is the capital of Jiangsu Province, located in the lower Yangtze River Plain of eastern China, and serves as a major economic center in the Yangtze River Delta. Existing studies suggest that PM_{2.5} in Nanjing primarily originates from industrial-related emissions, in which coal combustion and heavy-oil combustion are prominent sources [24]. An official report

further indicates that the two steel enterprises in Nanjing produce nearly 18 million tons of crude steel annually, and their atmospheric pollutant emissions account for more than 60% of the city's industrial-source emissions. Coal and heavy oil are the main industrial fuels for industries such as steelmaking. In 2020, the total coal consumption in Nanjing reached 29.9681 million metric tons, exceeding the average level of cities in the Yangtze River Delta. Coal burning emits primary particles and releases gaseous precursors such as SO_2 and NO_x , which facilitate secondary aerosol formation and elevate $\text{PM}_{2.5}$ concentrations. This industrial structure and energy-use profile subject Nanjing to sustained $\text{PM}_{2.5}$ emission pressures.

The city's topography, surrounded by mountains on three sides and water on one, also limits horizontal and vertical diffusion of pollutants, especially during stagnant meteorological conditions, which often leads to severe haze episodes [25]. Four national monitoring stations—Pukou (1157A), Gulou (1151A), Qixia (1159A), and Jianye (1158A)—are strategically located to represent industrial, urban, and residential areas across the city (Figure 1). In 2020, the annual mean $\text{PM}_{2.5}$ concentration in Nanjing reached $39 \mu\text{g}/\text{m}^3$, still exceeding China's national air quality standard ($35 \mu\text{g}/\text{m}^3$), reflecting persistent air quality challenges.

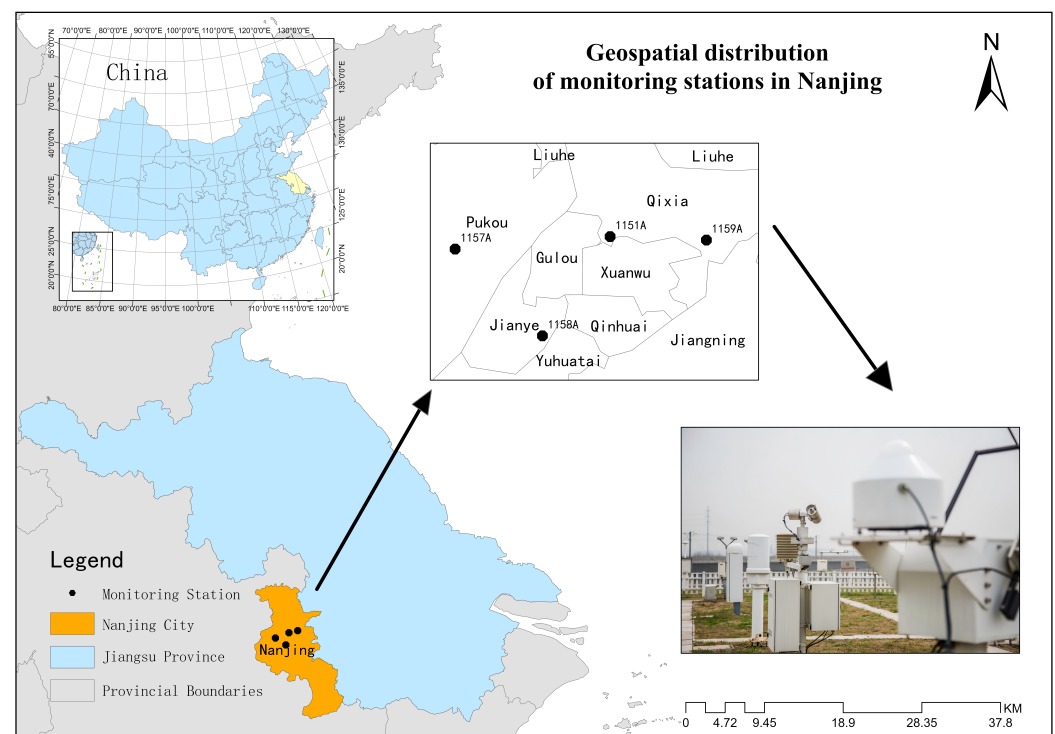


Figure 1. Geospatial distribution of monitoring stations in Nanjing.

Historical pollution events, such as those in December 2017 and January 2020 [26], revealed that sudden increases in particulate concentrations are often poorly captured by traditional deterministic models, resulting in missed or false warnings. In this context, there is an urgent need for uncertainty-aware forecasting frameworks capable of capturing rapid pollution transitions. This study introduces an FBA-LSTM model to address these challenges by integrating spatiotemporal features and probabilistic prediction, thereby improving the reliability of air quality forecasts and public health warnings in Nanjing.

2.2. Methodology

2.2.1. Model Overview and Design Rationale

The proposed FBA-LSTM model is a hybrid deep learning framework that integrates Bayesian LSTM, residual connections, and Bahdanau's attention mechanism to enhance both the prediction accuracy and uncertainty quantification in $PM_{2.5}$ forecasting (Figure 2). The design aims to capture complex temporal dependencies while ensuring stable training and reliable predictive confidence.

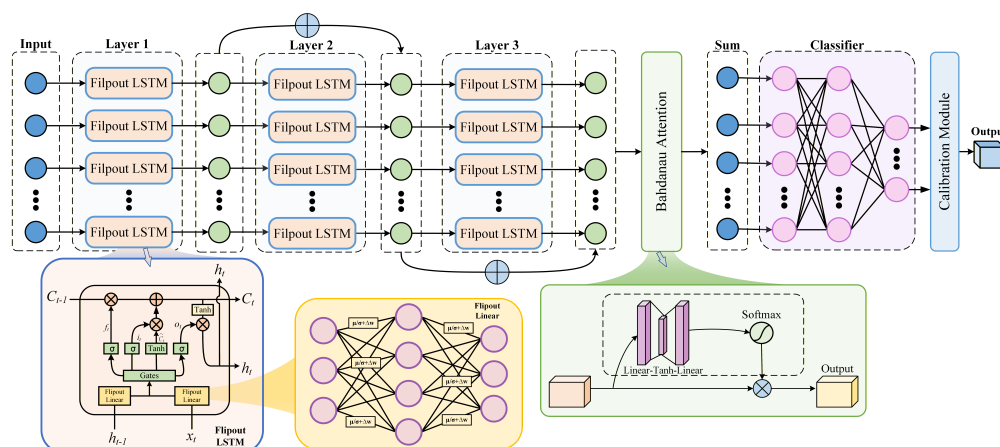


Figure 2. Architecture of the FBA-LSTM model in this study: overall flowchart of the model; internal structure of the flipout LSTM block; and internal structure of the flipout linear block.

At its core, the model employs a three-layer Bayesian LSTM, where each layer utilizes the flipout inference method for efficient Bayesian learning. The Bayesian LSTM incorporates flipout-based linear transformations and the reparameterization technique to model weight uncertainty. In each layer, the flipout linear module receives the current input X_t and the previous hidden state H_{t-1} , and generates the information required for the four LSTM gates: the forget gate (F_t), input gate (I_t), cell state (C_t), and output gate (O_t). Each weight parameter is decomposed into a mean (μ) and a standard deviation (α), and random perturbations (δW) are sampled from a standard normal distribution during forward propagation. This stochastic perturbation mechanism enables the model to learn a distribution over weights, thereby improving uncertainty estimations and reducing the gradient variance.

To alleviate vanishing or exploding gradients in deep recurrent architectures, residual connections were introduced after each Bayesian LSTM layer. These identity skip links maintain the flow of information across layers, enhancing both the stability and convergence efficiency. Furthermore, a Bahdanau attention mechanism is applied at the output stage to capture long-term temporal dependencies. By computing attention weights across a 24-h time window, the model focuses on critical time steps such as abrupt pollution events or seasonal transitions, thereby improving the interpretability and prediction accuracy.

Finally, a post hoc calibration module was added to refine the predictive confidence. Three calibration strategies—standard deviation calibration, interval calibration, and isotonic regression calibration—were compared, and the best-performing method was selected to align the predicted uncertainty with empirical data. This adjustment reduces overconfidence and produces well-calibrated uncertainty estimates, enhancing the calibration quality and utility for downstream warning tasks.

To clarify the correspondence between the variables discussed in Section 2.3 and the data presented in Section 2.2, and to facilitate the understanding of subsequent model formulas, Table 1 summarizes the role of these variables within the FBA-LSTM model

pipeline. The table provides comprehensive mapping, explaining the meaning, data source, and corresponding role of each variable in the model architecture. It focuses on illustrating the processing of various variables, such as the raw feature vector z_t , the model input sequence X_t , and the LSTM outputs H_t and C_t .

Table 1. Consistency between the variables in Section 2.2 and the research data in Section 2.3, and their corresponding relationships in the FBA-LSTM.

Variables	Meaning	Data Source in Section 2.2	Correspondence in the Model
$z_t \in \mathbb{R}^d$	Raw feature vector at time t (before entering the network)	Final feature set after selection; d is the number of selected features	Standardized input features after missing-value handling and Min–Max normalization
X_t	Model input sequence at time t	Constructed from normalized z_t via a 24-h sliding window	Input to the stacked three-layer Bayesian LSTM
H_t	Hidden-state output of the LSTM layer at time t	Computed by the LSTM layer (internal state)	Input to the next layer
C_t	Cell state of the LSTM layer at time t	Computed by the LSTM layer (internal state)	Used in LSTM for internal recurrence, to calculate H_t
x_l	Residual-block input at layer index l	Derived from the previous LSTM layer output	Equal to the LSTM output sequence ($x_l \equiv H_t$)
$F(x_l, W_l)$	Nonlinear transformation on the residual branch	Corresponds to the next-layer LSTM mapping	Raw output of the next LSTM before residual addition
$x_{l+1} = x_l + F(x_l, W_l)$	Residual-summed output	Element-wise summation of the previous two terms	Input to the next LSTM layer
$\Delta W_n, (\mu, \alpha)$	Flipout weight perturbation and parameters	Model parameters	Applied to LSTM weights for stochastic H_t and uncertainty estimation
$\alpha_{t,s}$	Attention weight for time step s contributing to output at t	Similarity scores from the last-layer output sequence	Computed by Bahdanau attention to weight and aggregate temporal information
C_t^{att}	Attention context vector	Weighted sum of the last-layer hidden states using $\alpha_{t,s}$	Passed to the linear output layer to generate the final prediction
$\hat{y}_t$	Predicted PM _{2.5} value	Corresponds to the ground-truth PM _{2.5} in Section 2.2	Produced by the linear layer; Bayesian sampling yields a predictive distribution and confidence intervals

t denotes the time index, whereas l denotes the layer index in the residual formulation. We use C_t^{att} for the attention context vector and reserve C_t for the LSTM cell state to avoid ambiguity.

2.2.2. Bayesian LSTM with Flipout Inference

We adopted an LSTM as the backbone sequence model to capture temporal dependencies in PM_{2.5} dynamics, leveraging its gated memory units to retain long-term information and alleviate gradient vanishing during sequence learning [27]. The LSTM updates its internal memory by combining the candidate cell state with the outputs of the input and forget gates, thereby regulating the integration of new information and the retention of previous context. The output gate further modulates the transformed cell state to generate the hidden representation at the current time step.

$$\begin{aligned}
 \tilde{C}_t &= \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \\
 C_t &= F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \\
 H_t &= O_t \odot \tanh(C_t)
 \end{aligned} \tag{1}$$

The formulas for the input gate I_t , forget gate F_t , and output gate O_t are as follows:

$$\begin{aligned} I_t &= \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \\ F_t &= \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \\ O_t &= \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \end{aligned} \quad (2)$$

where σ and $\tanh$ are two nonlinear activation functions; W_{xi} , W_{xf} , W_{xo} , W_{hi} , W_{hf} , and W_{ho} are weight parameters; and b_i , b_f , b_o are bias parameters.

To enable principled uncertainty quantification in the LSTM, we adopted a Bayesian treatment of the model parameters, viewing them as random variables rather than fixed quantities. This probabilistic formulation captures epistemic uncertainty by inferring the posterior distribution over parameters conditional on the observations. Its formula is as follows:

$$p(\theta | y_0) = \frac{p(y_0 | \mathcal{D}) p(\theta)}{p(y_0)} \quad (3)$$

where θ is the model parameter, y_0 is the observed data, $p(\theta | y_0)$ is the posterior distribution, $p(y_0 | \mathcal{D})$ is the likelihood function, $p(\theta)$ is the prior distribution, and $p(y_0)$ is the marginal likelihood.

Exact posterior inference in deep neural architectures is analytically intractable due to the high dimensionality and nonlinearity of the parameter space. We therefore adopted variational inference (VI) as an efficient approximation. VI recasts posterior inference as an optimization problem by introducing a tractable variational distribution $q(\theta | \lambda)$ to approximate the true posterior $p(\theta | y_0)$. The objective is to minimize the discrepancy between these distributions, typically quantified by the Kullback–Leibler (KL) divergence; its formula is as follows:

$$\text{KL}(q(\theta | \lambda) \| p(\theta | y_0)) = -\mathcal{L}(\lambda) + \log p(y_0) \quad (4)$$

Because evaluating the marginal likelihood $p(y_0)$ is intractable when minimizing the KL divergence, variational inference instead maximizes the evidence lower bound (ELBO), expressed as:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\theta|\lambda)}[\log p(y_0, \theta) - \log q(\theta | \lambda)] \quad (5)$$

where $\mathbb{E}_{q(\theta|\lambda)}$ denotes the expectation of the variational distribution $q(\theta | \lambda)$. $\mathcal{L}(\lambda)$ is the ELBO of λ , and by maximizing ELBO, the KL divergence between the variational distribution and the true posterior distribution can be minimized to obtain the optimal posterior approximation.

Finally, we adopted flipout to generate per-example pseudo-independent weight perturbations within each minibatch, decorrelating gradient estimates while avoiding redundant forward passes, thereby stabilizing and accelerating the training of the Bayesian LSTM.

$$\Delta W_n = \hat{W} \odot r_n s_n^T \quad (6)$$

where W_n is the perturbation of the n -th sample; $\hat{W}$ is the shared baseline weight perturbation; and r_n and s_n are independently sampled random vectors from a Bernoulli distribution of ± 1 .

In this way, the flipout method ensures that the gradient of each sample is independent of other samples, thereby effectively reducing the variance of the gradient estimate. In flipout's neural network forward propagation, the weights of each sample are updated in the following way:

$$y_n = f\left(\left(W + \hat{W} \odot r_n s_n^T\right) x_n\right) \quad (7)$$

where y_n is the output of the n -th sample and W is the basic weight matrix in the network.

2.2.3. Attention Mechanism

Within the Bayesian LSTM, we adopted the Bahdanau attention mechanism to improve the model performance by aligning the decoder with the most relevant encoder representations [28]. At each decoding step t , a feed-forward scoring function measures the compatibility between the decoder's hidden state H_t and each encoder's hidden state H_s , producing unnormalized scores $e_{t,s}$:

$$e_{t,s} = v_a^\top \tanh(W_h H_s + W_t H_t + b_a), \quad (8)$$

where W_h and W_t are learnable weight matrices, v_a and b_a are learnable parameters, and $\tanh(\cdot)$ is the activation function. The scores $e_{t,s}$ are normalized via a softmax to obtain the attention weights $\alpha_{t,s}$:

$$\alpha_{t,s} = \frac{\exp(e_{t,s})}{\sum_{j=1}^T \exp(e_{t,j})}. \quad (9)$$

The weight $\alpha_{t,s}$ quantifies the contribution of encoder position s to the decoder at time t . The context vector is computed as

$$C_t^{\text{att}} = \sum_{s=1}^T \alpha_{t,s} H_s, \quad (10)$$

Subsequently, the context vector C_t^{att} is combined with the current decoder state to produce the final output.

2.2.4. Residual Connections

To enhance gradient propagation and stabilize the optimization of deep neural architectures, this study adopted a residual connections strategy. Let x_l denote the input feature of the l -th layer, $F(\cdot, W_l)$ represent a nonlinear transformation parameterized by W_l , $h(\cdot)$ define the identity or projection mapping on the shortcut path, and $f(\cdot)$ specify an optional transformation applied after the residual summation, typically an activation or normalization operation.

$$y_l = h(x_l) + F(x_l, W_l), \quad x_{l+1} = f(y_l). \quad (11)$$

With in the Bayesian LSTM, the shortcut connection is implemented as identity mapping, and no additional nonlinearity is applied after the summation. Since all layers share the same feature dimension, the residual addition is element-wise and parameter-free:

$$x_{l+1} = x_l + F(x_l, W_l). \quad (12)$$

For a network with multiple stacked residual blocks, the relationship between a deeper layer L and an earlier layer l can be written as:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i), \quad (13)$$

This structure facilitates effective information preservation and gradient flow throughout the network, enabling stable training of deep recurrent or attention-based architectures.

2.2.5. Model Calibration Methods and Training Settings

While the rapid advancement of deep learning has markedly improved model accuracy, it has also reduced predictive confidence [23]. To enhance the reliability of predictive uncertainty estimates, a calibration procedure was applied to the post-trained FBA-LSTM model. The purpose of model calibration is to align predicted outputs with observed values, thereby producing well-calibrated confidence estimates. In this study, three representative post hoc calibration methods—standard deviation calibration, interval calibration, and isotonic regression calibration—were systematically evaluated [29–31], and the most effective approach was ultimately adopted. Their theoretical principles are summarized as follows:

- (1) Standard deviation calibration: Rescales the predictive variance so that the spread of predictions matches the residual variability observed on the calibration set, aligning the predicted uncertainty with actual errors.
- (2) Interval calibration: Adjusts prediction intervals so their empirical coverage matches the nominal level, correcting under- or over-confident intervals.
- (3) Isotonic regression calibration: Learns a monotonic, non-parametric mapping from predicted confidence to observed coverage, aligning predicted probabilities with empirical frequencies.

For the proposed FBA-LSTM, we stacked three flipout LSTM layers, each with 50 hidden units. The model's loss function comprises MSE and KL divergence. In Stage I, β is fixed to 0 and the model is trained only under the MSE objective. When the improvement in MSE over the most recent five epochs is smaller than 0.3, the training is considered to have converged and switches to Stage II. In Stage II, β is set to 0.3 times the MSE value divided by the KL value and is updated proportionally within each mini-batch, ensuring that the weighted KL term contributes no more than 30% of the total loss. Inputs use a 24-h sliding window, and the Bahdanau attention layer has 50 units. Optimization uses AdamW with an initial learning rate of 1×10^{-3} , a batch size of 32, and 150 training epochs. The baseline models residual LSTM, attention LSTM, and MC dropout LSTM share the same backbone of three LSTM layers with 50 units per layer and are trained with AdamW under an MSE loss; they use the same 24-h window, a batch size of 32, and 150 training epochs. The residual variant introduces identity skip connections, the attention variant includes a temporal attention module, and the MC dropout variant enables stochastic predictions by keeping dropout active during training and inference.

2.3. Data Collection and Processing

2.3.1. Data Acquisition

To comprehensively assess the proposed model, this study employed hourly air quality and meteorological datasets from Nanjing, China. Air pollution and meteorological factors jointly govern the formation, diffusion, and transformation of fine particulate matter (PM_{2.5}); integrating these datasets offers a robust and scientifically grounded basis for PM_{2.5} concentration predictions. The air pollution dataset was acquired from the real-time monitoring system of the China Environmental Monitoring Station (<https://www.cnemc.cn>). It includes seven key indicators: the air quality index (AQI), carbon monoxide (CO), nitrogen dioxide (NO₂), ozone (O₃), fine particulate matter (PM_{2.5}), inhalable particulate matter (PM₁₀), and sulfur dioxide (SO₂). These pollutants are closely associated with PM_{2.5} variations because of their shared anthropogenic emission sources and their participation in photochemical reactions within the atmosphere. The meteorological dataset was obtained from the U.S. National Oceanic and Atmospheric Administration (NOAA) (<https://www.noaa.gov>). It contains seven variables—temperature, dew point temperature, barometric pressure, wind speed, wind direction, cloud cover, and six-hour cumulative precipitation—

that substantially influence pollutant dispersion, chemical transformation, and secondary aerosol formation processes.

Both datasets spanned the period from 1 January to 31 December 2020, with an hourly temporal resolution, resulting in a total of 8784 observations. For model construction and evaluation, the data were partitioned into training (60%), validation (20%), and testing (20%) subsets, corresponding to 5263, 1755, and 1755 hourly samples, respectively. The division was performed in a chronological manner, where earlier data were used for model training, followed sequentially by validation and testing. This strategy preserves the temporal continuity of atmospheric processes and effectively prevents information leakage across time. Table 2 shows the statistical analysis of the meteorological and air pollution datasets.

Table 2. Statistical results of meteorological and air pollution datasets.

Variables	Abbreviation	Unit	Max	Min	Mean
PM _{2.5}	–	µg/m ³	172.0	1.0	26.0
Air Quality Index	AQI	–	347.0	8.0	54.6
Carbon Monoxide	CO	ppm	2.7	0.1	0.8
Nitrogen Dioxide	NO ₂	µg/m ³	151.0	1.0	31.3
Ozone	O ₃	µg/m ³	318.0	1.0	70.4
PM ₁₀	–	µg/m ³	457.0	1.0	57.5
Sulfur Dioxide	SO ₂	µg/m ³	227.0	1.0	7.3
Temperature	Temp	°C	38.0	−10.0	17.0
Dew Point Temperature	DPT	°C	30.2	−16.6	12.3
Sea Level Air Pressure	SLP	hPa	1040.8	997.1	1017.0
Wind Direction	WD	°	360.0	0.0	154.43
Wind Speed	WS	m/s	9.0	0.0	2.5
Cloud Cover	CC	–	9.0	0.0	4.8
6-Hour Rainfall	6 h Rain	mm	72.0	0.0	0.9

2.3.2. Data Processing

The acquired dataset contained some missing values. To ensure data completeness and accuracy, it is necessary to adopt effective methods for handling these missing values. Depending on the nature of different features, we applied various interpolation techniques. For features with relatively smooth and linear variations (e.g., wind speed and cloud cover), we used linear interpolation. For features with significant fluctuations and nonlinear variations (e.g., PM_{2.5}, AQI, CO, NO₂, and other air pollutant concentrations), cubic spline interpolation was employed. The wind direction, expressed as angular values between 0° and 360°, exhibits a natural periodicity that traditional interpolation methods cannot effectively capture. To preserve this cyclic property, we interpolated the sine and cosine components of the wind direction and then converted the results back to angles, ensuring physical consistency and preventing outliers. Finally, we applied the Max–Min normalization method to process all input features, eliminating the differences in magnitudes between features, ensuring balanced contributions during model training, and accelerating convergence while improving the model stability.

The technical workflow of this study, illustrated in Figure 3, consisted of data acquisition, data preprocessing, model training and testing, and finally, model evaluation. In the data acquisition stage, air pollution data and meteorological data were obtained from the China National Environmental Monitoring Center and the National Oceanic and Atmospheric Administration, respectively. The next stage was data preprocessing, where 12 candidate features related to PM_{2.5} formation and variation were selected based on domain knowledge and prior research. The features were then filtered by their Pearson correlation, resulting in the final selection of the feature set most strongly correlated with

PM_{2.5}. Afterward, missing values in the selected features were handled using different interpolation methods, and Min–Max normalization was applied to standardize the features, ensuring consistency between them and preventing discrepancies in the scale that could affect model training. After preprocessing, the data were fed into the model training and testing stage, where the Flipout LSTM model was trained on the processed data. In the post-processing stage, the model was calibrated to output reliable predictions. Finally, in the model evaluation stage, the model’s predictive accuracy and reliability were assessed using various metrics, such as the RMSE, MAE, and F1 score.

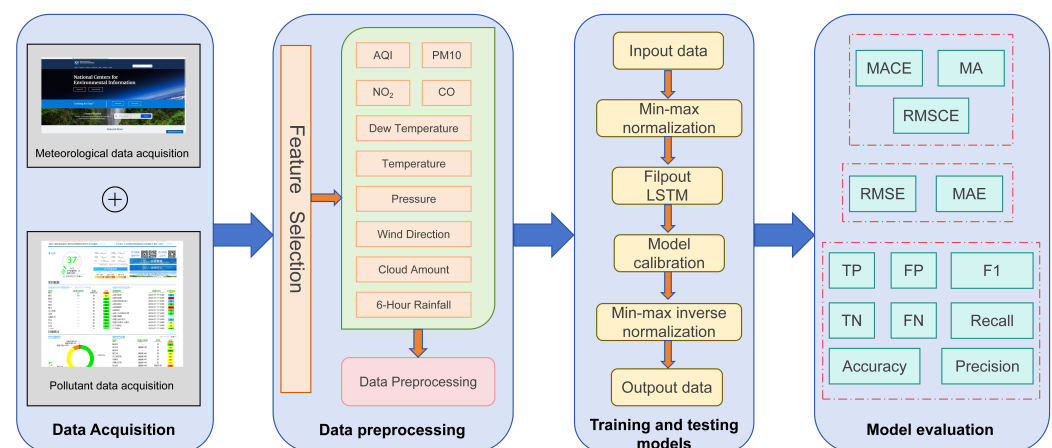


Figure 3. Technical workflow diagram for this study.

2.3.3. Feature Selection

Thirteen candidate features were selected based on established atmospheric science knowledge and prior studies highlighting their strong associations with PM_{2.5} formation and variation [14,32]. These features included the AQI, PM₁₀, NO₂, CO, dew point temperature, temperature, barometric pressure, wind direction, cloud cover, six-hour precipitation, O₃, wind speed, and SO₂. A correlation analysis was then performed to evaluate the relationships between each variable and the PM_{2.5} concentration. Consequently, ten representative features exhibiting strong correlations with PM_{2.5} were retained for model development, while three features were excluded due to weak associations.

To evaluate the correlation between each feature and the PM_{2.5} concentration, we calculated the absolute value of the Pearson correlation coefficient (PCC). The Pearson correlation coefficient measures the strength and direction of a relationship between two variables, ranging from −1 to +1. The formula is:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (14)$$

where r is the Pearson correlation coefficient; X_i and Y_i represent the values of the i -th data point for variables X and Y , respectively; $\bar{X}$ and $\bar{Y}$ are the sample means of variables X and Y ; and n is the sample size.

The results of the correlation analysis are presented in Figure 4a,b. The AQI, PM₁₀, NO₂, CO, dew point temperature, air temperature, and barometric pressure exhibited strong positive correlations with PM_{2.5} concentrations. The cumulative contribution of the top 10 features exceeded 99%, indicating that most of the variance in PM_{2.5} can be captured by these indicators. In contrast, SO₂, the wind speed and O₃ showed extremely low weight coefficients (all below 0.005), with a combined contribution of less than 0.5%. Therefore, these features were excluded from the subsequent modeling process to eliminate

information redundancy and enhance both the computational efficiency and the model prediction performance.

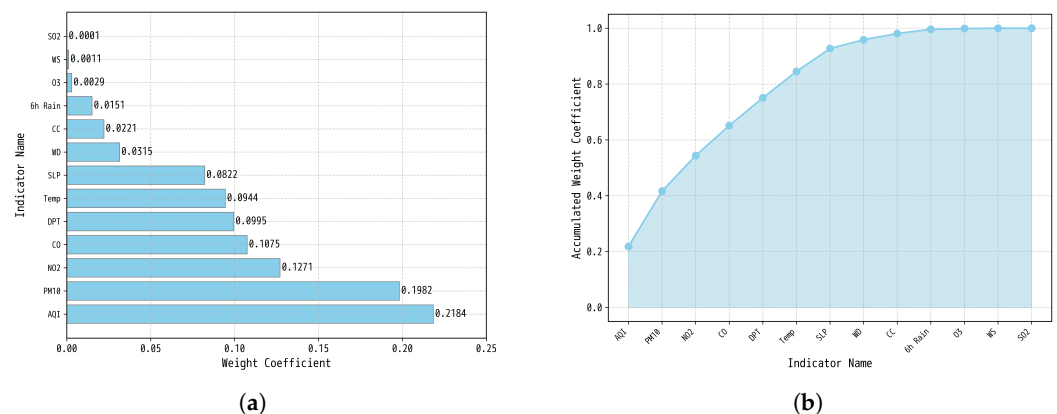


Figure 4. Indicator importance and cumulative contribution for PM_{2.5} predictions. (a) Bar chart of indicator weights. (b) Cumulative feature weights.

2.4. Evaluation Metrics

To evaluate the performance of the calibration methods, we employed three metrics—MACE, RMSCE, and MA. The calculation formulas are as follows:

$$\text{MACE} = \frac{1}{K} \sum_{k=1}^K |\hat{p}_k - p_k| \quad (15)$$

$$\text{RMSCE} = \sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{p}_k - p_k)^2} \quad (16)$$

$$\text{MA} = \int_0^1 |\hat{p}(x) - p(x)| dx \quad (17)$$

where K is the number of prediction intervals; $\hat{p}_k$ is the predicted coverage and p_k is the observed coverage; and $\hat{p}(x)$ and $p(x)$ represent the predicted and observed coverage as continuous functions over confidence level $x \in (0, 1)$.

To evaluate the performance of different models in the PM_{2.5} prediction task, we employed two metrics—RMSE and MAE. The calculation formulas are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (18)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (19)$$

where n is the sample size, y_i is the true value of the i -th sample, and $\hat{y}_i$ is the predicted value for the i -th sample.

For threshold-based warning evaluations, the predictions were binarized at 35 $\mu\text{g}/\text{m}^3$ (24-h PM_{2.5}), which corresponds to the upper limit of the excellent air quality level according to the Technical Regulation on Ambient Air Quality Index (HJ 633-2012) [33]. We constructed a confusion matrix with: TPs (true positives), TNs (true negatives), FPs (false positives), and FNs (false negatives). The accuracy, precision, recall, and F1 score were then computed from this matrix to assess the exceedance-detection performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (20)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (21)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (22)$$

$$F1 = \frac{2 TP}{2 TP + FP + FN} \quad (23)$$

3. Results

3.1. Comparison of Model Calibration Methods

In the post-processing calibration stage of the FBA-LSTM model, three different calibration methods were compared, and their performance was quantitatively evaluated using three metrics: the mean absolute calibration error (MACE), the root mean square calibration error (RMSCE), and the miscalibration area (MA). The experimental results are presented in Table 3 and Figure 5.

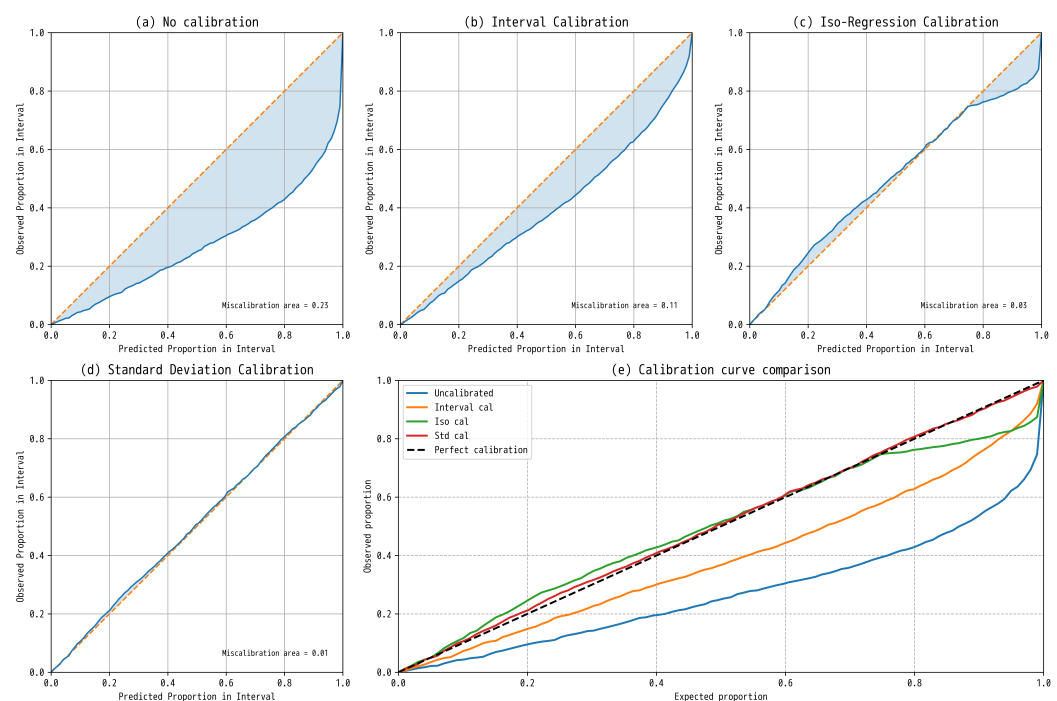


Figure 5. Comparison of different calibration methods. (a) No calibration: calibration curve without any calibration processing. (b) Interval calibration: calibration curve under interval calibration. (c) Iso-regression calibration: calibration curve using iso-regression calibration. (d) Standard deviation calibration: calibration curve with standard deviation calibration. (e) Calibration curve comparison: comparative curve of different calibration methods (including the perfect calibration baseline).

Table 3. Post-processing calibration performance evaluation.

Method	MACE	RMSCE	MA
No Calibration	0.22521	0.25382	0.22748
Interval Calibration	0.10620	0.11924	0.10728
Iso-Regression Calibration	0.03446	0.04840	0.03480
Standard Deviation Calibration	0.00740	0.00880	0.00740

In the uncalibrated case, the model exhibited an MACE of 0.22521, an RMSCE of 0.25382, and an MA of 0.22748. The significant deviation of the uncalibrated curve (Figure 5a) from the ideal calibration line indicates a systematic discrepancy between the predicted confidence and the actual observed proportions, suggesting that the model suffers from a degree of overconfidence.

After applying the interval calibration method, the MACE, RMSCE, and MA of the model were reduced to 0.10620, 0.11924, and 0.10728, respectively. The calibration curves (Figure 5b) showed improvement compared to the uncalibrated case, with the error region narrowing to 0.11. However, the overall performance remained higher than the other two methods, indicating that this method has limited effectiveness in reducing calibration errors.

The isotonic regression calibration method produced favorable results at low to medium confidence intervals (below approximately 64%). The MACE, RMSCE, and MA of this method were 0.03446, 0.04840, and 0.03480, respectively, outperforming the interval calibration method. However, its performance declined at higher confidence intervals. The error region of the calibration curve (Figure 5c) was 0.03, which was lower than that of the interval calibration method, but remained relatively high.

The standard deviation calibration method achieved the best results across all three evaluation metrics, with an MACE of 0.00740, an RMSCE of 0.00880, and an MA of 0.00740. The calibration curves (Figure 5d) closely align with the ideal calibration line, with an error region of only 0.01, demonstrating a stable and consistent calibration performance.

Among the three approaches, the standard deviation calibration yielded the lowest calibration errors (MACE, RMSCE, MA; see Table 3) and the smallest deviation from the ideal line (Figure 5d, error region ≈ 0.01). Compared with interval calibration (Figure 5b), which improved the confidence to some extent, but still exhibited larger errors across confidence levels, and isotonic regression (Figure 5c), which performs well at a low–medium confidence, but degrades at higher levels, the standard deviation calibration method works by rescaling the predictive variance, correcting the model’s overall dispersion bias and ensuring a more consistent alignment across all confidence levels, addressing the misalignment issue observed in high-confidence intervals with isotonic regression calibration. This produces a more consistent alignment across the full confidence range. Accordingly, we adopted standard deviation calibration as the default post-processing scheme for the FBA-LSTM.

3.2. Prediction Accuracy Comparison

Under the noiseless condition, the FBA-LSTM performed relatively well, achieving the lowest RMSE and MAE values in three out of four stations. For instance, at site 1157A, the RMSE and MAE of the FBA-LSTM were 6.39 and 4.77, respectively, which were lower than those of the standard LSTM (RMSE of 7.26 and MAE of 5.40). The residual LSTM and attention LSTM also reduced the errors relative to LSTM, but remained slightly inferior to FBA-LSTM (Figure 6a), and the same pattern held at sites 1158A and 1159A (Figure 6c,d). At site 1151A, the RMSE of the MC dropout LSTM was 7.53 and the MAE was 5.12, slightly lower than the FBA-LSTM values (RMSE of 7.64 and MAE of 5.15) (Figure 6b).

Under noise conditions, Gaussian noise was added to the test data to simulate real-world measurement errors or data uncertainties. The noise strength was set to 10% of the data’s standard deviation. Under this setting, the prediction performance of all models decreased, but the extent of degradation varied across models. As point prediction models, the LSTM, residual LSTM, and attention LSTM showed a more noticeable performance degradation under noise interference. For example, at site 1159A, the RMSEs of the LSTM, residual LSTM, and attention LSTM increased by 1.76, 0.78, and 1.36, respectively, while the

MAEs increased to 1.40, 0.66, and 1.00 (Figure 6d). In contrast, the uncertainty prediction models, the MC dropout LSTM and FBA-LSTM, showed less performance degradation under noise conditions. The RMSE and MAE increases at the four monitoring stations did not exceed 1, demonstrating a better robustness. These findings demonstrate that uncertainty-aware models are more resilient to data noise, which is critical for reliable air-quality forecasting. Among these two models, the FBA-LSTM consistently outperformed the MC dropout LSTM. For example, at site 1157A, the difference between the two models was minimal in the absence of noise (RMSE of 6.39 vs. 6.49; MAE of 4.77 vs. 4.78). However, when noise was present, the FBA-LSTM performed better, with a lower RMSE and MAE (7.05 vs. 7.27; 5.38 vs. 5.52) (Figure 6b). This difference may be attributed to their respective uncertainty modeling approaches: the MC dropout LSTM performs approximate Bayesian inference by randomly dropping neurons, while the FBA-LSTM adopts the flipout inference method, which directly samples the weight distributions, offering a more comprehensive reflection of parameter uncertainty.

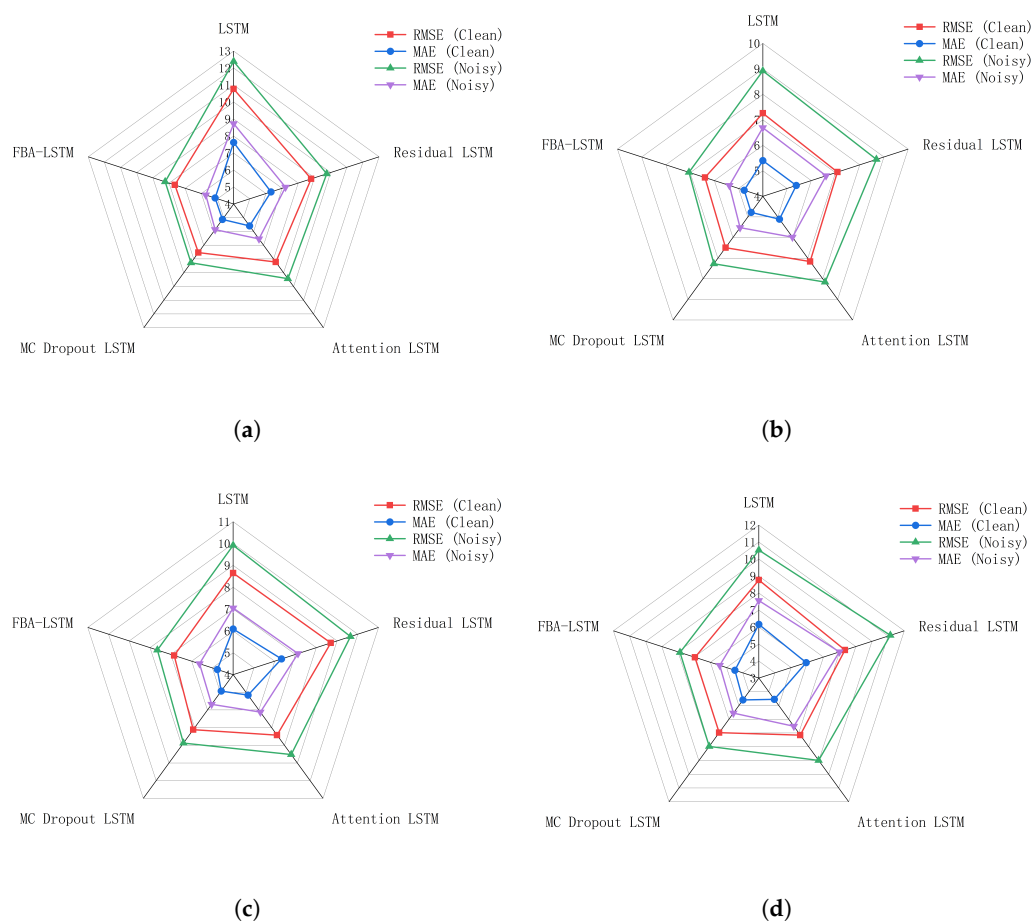


Figure 6. Performance comparison radar charts of LSTM-based models under clean and noisy conditions across four monitoring stations. (a) Radar chart for monitoring station 1151A. (b) Radar chart for monitoring station 1157A. (c) Radar chart for monitoring station 1158A. (d) Radar chart for monitoring station 1159A.

Considering the results across the four monitoring stations, the FBA-LSTM maintained low RMSE and MAE values both with and without noise, with minimal performance degradation under noise interference. This indicates that the FBA-LSTM exhibits a better robustness and generalization ability in environments with substantial observation errors (Table 4).

Table 4. Predictive performance of LSTM-based models evaluated under clean and noisy conditions.

Station	Model Type	No-Noise Condition		Noise Condition	
		RMSE	MAE	RMSE	MAE
1151A	LSTM	10.77	7.62	12.39	8.73
	Residual LSTM	8.80	6.31	9.76	7.21
	Attention LSTM	8.22	5.59	9.43	6.54
	MC Dropout LSTM	7.53	5.12	8.27	5.86
	FBA-LSTM	7.64	5.15	8.24	5.72
1157A	LSTM	7.26	5.40	8.92	6.69
	Residual LSTM	7.08	5.38	8.69	6.61
	Attention LSTM	7.16	5.11	8.16	5.98
	MC Dropout LSTM	6.49	4.78	7.27	5.52
	FBA-LSTM	6.39	4.77	7.05	5.38
1158A	LSTM	8.65	6.09	9.92	7.05
	Residual LSTM	8.70	6.33	9.65	7.12
	Attention LSTM	7.42	5.16	8.52	6.12
	MC Dropout LSTM	7.11	4.93	7.87	5.68
	FBA-LSTM	6.84	4.76	7.65	5.62
1159A	LSTM	8.77	6.16	10.53	7.56
	Residual LSTM	8.33	5.93	11.14	7.99
	Attention LSTM	7.15	4.55	8.99	6.51
	MC Dropout LSTM	6.97	4.59	7.97	5.55
	FBA-LSTM	6.95	4.48	7.88	5.42

Bold values denote the minimum RMSE and MAE for each monitoring station, indicating the best predictive performance of LSTM-based models for PM_{2.5} at this station under the corresponding conditions (no-noise/noise).

3.3. Early Warning Performance

According to the Technical Regulation on Ambient Air Quality Index (on trial) [33] (Table 5), PM_{2.5} concentrations are categorized into six levels. In this regulation, the PM_{2.5} breakpoints used for level classification are defined based on the 24 h average concentration for daily reporting.

Table 5. PM_{2.5} levels.

Index	PM _{2.5} (µg/m ³)
Excellent	0–35
Good	36–75
Light Pollution	76–115
Moderate Pollution	116–150
Heavy Pollution	151–250
Severe Pollution	>250

The air quality in Nanjing predominantly falls within the “Excellent”, “Good”, “Light Pollution”, and “Moderate Pollution” levels, with the majority of observations in the “Excellent” and “Good” categories (Figure 7). Accordingly, this study reformulated the PM_{2.5} concentration forecasting problem as a binary classification task for threshold exceedance detection. Specifically, the upper limit of the “Excellent” level, 35 µg/m³, was adopted as the classification threshold to identify time periods when pollution levels surpass the “Excellent” standard, corresponding to the occurrence of slight pollution and more severe conditions.

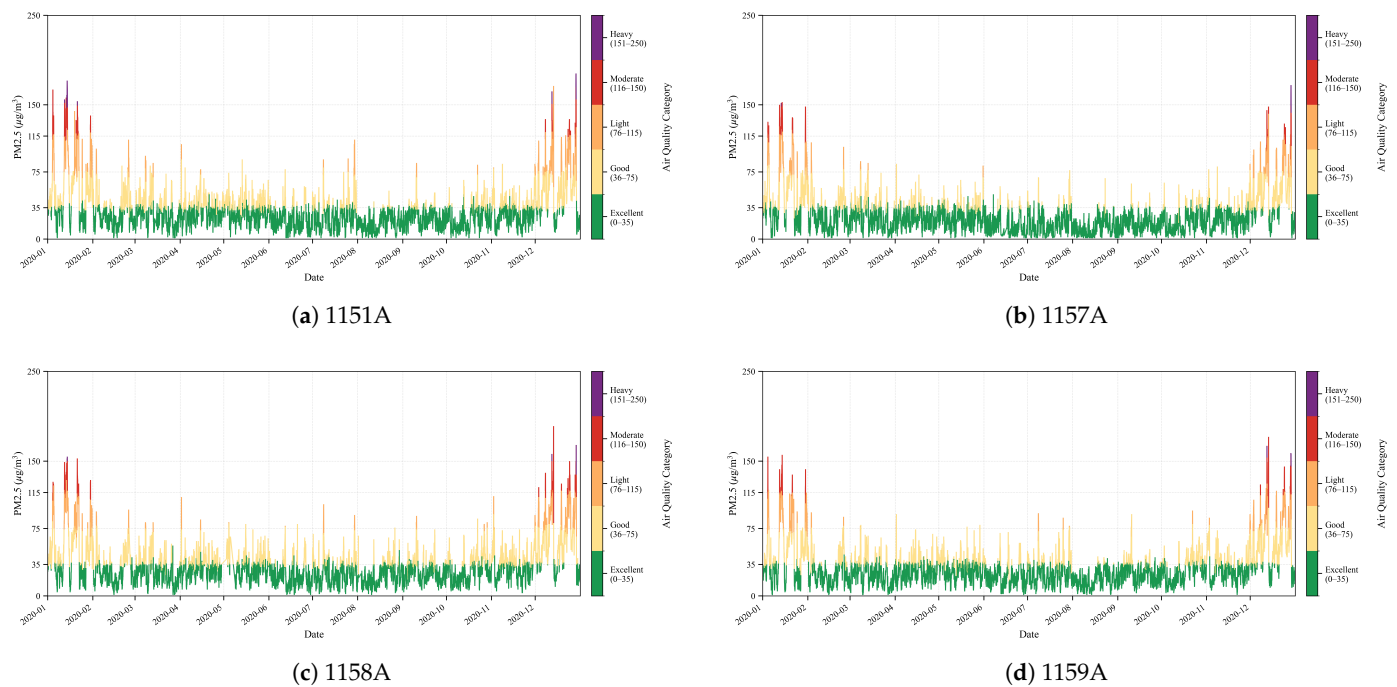


Figure 7. Annual PM_{2.5} concentration visualization at four monitoring stations, classified according to the HJ 633-2012 Air Quality Standard. Each panel displays the temporal variation in PM_{2.5} levels, with the color coding corresponding to the five-tier air quality index (excellent, good, light pollution, moderate pollution, heavy pollution).

As shown in Table 6, the five models differed in their performance on the PM_{2.5} threshold task. For deterministic models, the LSTM yielded TP = 620, TN = 1002, FP = 92, and FN = 38 (accuracy, 0.9258; precision, 0.8708; recall, 0.9422; F1 score, 0.9051). The residual LSTM reported TP = 610, TN = 1003, FP = 91, and FN = 48 (accuracy, 0.9207; precision, 0.8702; recall, 0.9271; F1 score, 0.8977). The attention LSTM showed a modest improvement with TP = 610, TN = 1021, FP = 73, and FN = 48 (accuracy, 0.9309; precision, 0.8931; recall, 0.9271; F1 score, 0.9098). All three models exhibited a limited discriminative ability around the critical threshold and were prone to making overconfident misjudgments, with a certain percentage of false positives and false negatives.

Table 6. Classification performance of different LSTM-based models.

Model Type	Accuracy	Precision	Recall	F1
LSTM	0.9258	0.8708	0.9422	0.9051
Residual LSTM	0.9207	0.8702	0.9271	0.8977
Attention LSTM	0.9309	0.8931	0.9271	0.9098
MC Dropout LSTM	0.9875	0.9941	0.9712	0.9825
FBA-LSTM	0.9966	0.9978	0.9934	0.9956

Compared to the aforementioned models, the MC dropout LSTM incorporates the Monte Carlo dropout method to quantify the prediction uncertainty through multiple forward propagations, identifying 309 high-uncertainty time periods (Figure 8a). Here, “high-uncertainty time periods” refers to timestamps at which the observed value lies outside the model’s 90% predictive interval, which is constructed as the predictive mean plus or minus the predictive standard deviation. The performance of the MC dropout LSTM model on the test set was as follows: 506 TPs, 919 TNs, 3 FPs, and 15 FNs, resulting in an accuracy of 0.9875, precision of 0.9941, recall of 0.9712, and F1 score of 0.9825. This model

achieved a better performance than the LSTM and attention LSTM in terms of the accuracy, precision, recall, and F1 score, with significantly fewer false positives and false negatives.

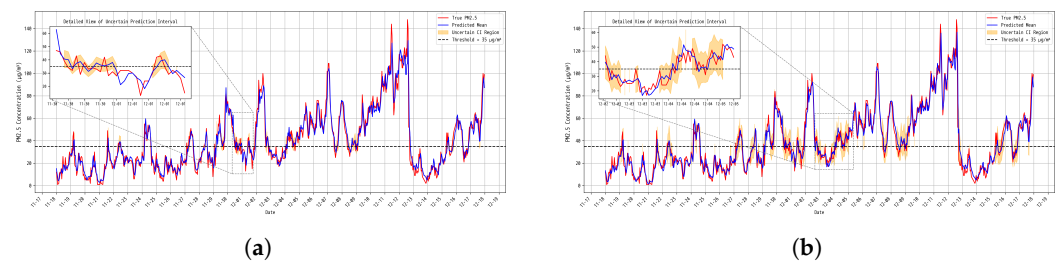


Figure 8. Visualization of prediction uncertainty intervals generated by MC dropout LSTM and FBA-LSTM. (a) Uncertain time periods identified by the MC dropout LSTM model (one-month interval). (b) Uncertain time periods identified by the FBA-LSTM model (one-month interval).

The FBA-LSTM model excelled in uncertainty quantification, identifying 567 high-uncertainty periods under the same decision criterion as the MC dropout LSTM (Figure 8b). Compared with dropout-based approximation methods, the flipout Bayesian inference model introduces probability distributions to model uncertainty in the network weights rather than relying on the heuristic operation of random deactivation, thereby more systematically capturing epistemic uncertainty in model predictions and showing a higher sensitivity to fluctuations near the critical threshold. The performance on the test set was as follows: TP = 454, TN = 727, FP = 1, and FN = 3, resulting in an accuracy of 0.9966, precision of 0.9978, recall of 0.9934, and F1 score of 0.9956. Compared to the other models, the FBA-LSTM achieved a false alarm rate of 0.14%, a miss rate of 0.66%, and classification performance metrics that were very close to one.

Overall, FBA-LSTM significantly reduces both false alarms and omissions while maintaining a high accuracy. Additionally, its ability to identify uncertain samples outperforms the other four models.

3.4. Outlier Detection

To evaluate the model's ability to identify anomalous events, this paper further designs a detection experiment based on manually injected anomalies. It compares and analyzes the FBA-LSTM method with the isolation forest (IF) and interquartile range (IQR) methods. We selected a month with few native outliers from the PM_{2.5} series and injected approximately 5% synthetic anomalies; the injection timestamps were drawn by random sampling without replacement, and the original values at those timestamps were scaled up by 2.5× or downscaled to 0.1× to emulate sudden pollution spikes and sensor failures. This setup allows for the evaluation of the performance of the different methods in the outlier detection task.

The anomaly detection results were quantitatively compared using three metrics: the accuracy, recall, and F1 score. These metrics were chosen to comprehensively assess the overall detection capability, the risk of omission, and the balance between accuracy and recall for each method in anomaly identification. This approach ensured that the evaluation results were both representative and interpretable.

As shown in Table 7, the FBA-LSTM achieved the highest scores across all three evaluation metrics, with an accuracy of 0.553, a recall of 0.722, and an F1 score of 0.627. These values were higher than those of the comparison methods, isolation forest (accuracy = 0.441, recall = 0.417, F1 = 0.217) and IQR (accuracy = 0.500, recall = 0.139, F1 = 0.217). The FBA-LSTM detects anomalies when an observation lies outside a high-confidence, heteroscedastic predictive interval inferred from the model's predictive mean and calibrated variance, allowing the band to widen in volatile regimes and narrow in stable periods. Unlike IF and

IQR, which use distributional heuristics without time-conditioned predictive uncertainty, this model-relative criterion reduces false alarms under normal variability while recovering more injected spikes and dropouts, consistent with the higher accuracy, recall, and F1.

Table 7. Performance comparison of anomaly detection methods.

Method	Accuracy	Recall	F1
IF	0.441	0.417	0.429
IQR	0.500	0.139	0.217
FBA-LSTM	0.553	0.722	0.627

Figure 9 further validates the above quantitative findings from a visual perspective. The FBA-LSTM successfully responded to multiple real injected anomalies without widespread misclassification, indicating that the model exhibits a high sensitivity and a low false alarm rate in anomaly detection. In contrast, the IQR method only detected a small number of anomalies with a large amplitudes and struggled to identify anomalies with slight deviations, resulting in a low recall rate. This is consistent with its reliance on statistical thresholds for global discrimination, which limits its ability to adapt to the dynamic nature of local data. On the other hand, while the IF method can detect significantly deviated outliers, it is less effective at recognizing mild outliers that are closer to the normal distribution. This limitation may stem from these anomalies not being sufficiently isolated in the feature space, failing to form a short enough path length during the randomization process, and thus not being effectively classified as anomalies.

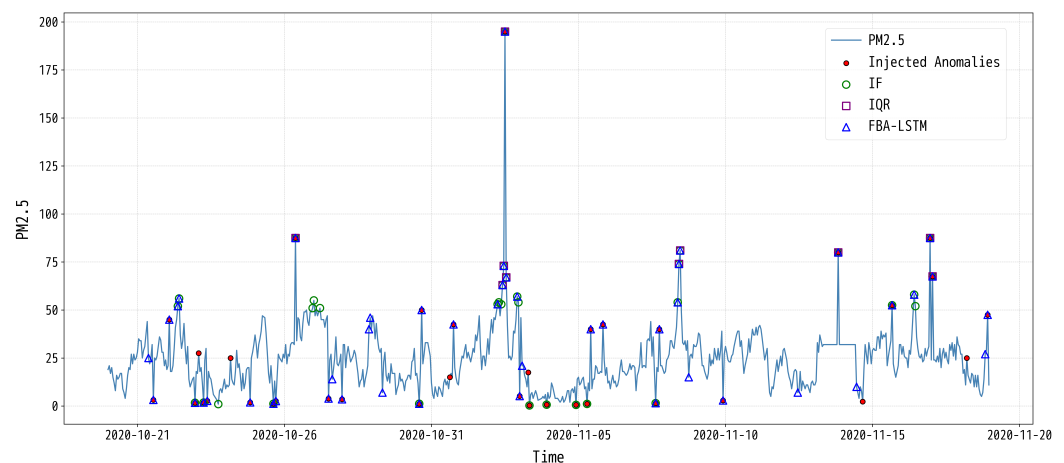


Figure 9. Comparative analysis of isolation forest, IQR, and FBA-LSTM for $PM_{2.5}$ anomaly detection (2020-10-20 to 2020-11-18).

By combining the quantitative results in Table 7 with the visual analysis in Figure 9, it is evident that the FBA-LSTM achieved a strong balance between accuracy, recall, and false alarm control, demonstrating a strong generalization ability and robustness.

4. Discussion

4.1. Discussion in the Context of Previous Studies

At present, a variety of studies have investigated LSTM-based frameworks for $PM_{2.5}$ forecasting, most of which have focused on improving the deterministic point-prediction accuracy. Chen et al. developed a PSO-LSTM model that uses particle swarm optimization to tune key LSTM hyperparameters for forecasting $PM_{2.5}$, PM_{10} , and O_3 , and reported an improved goodness-of-fit and reduced short-horizon errors compared with random forest and a plain LSTM [34]. Liu and Guo coupled an LSTM with the sparrow search algorithm

(LSTM-SSA) for multi-pollutant and AQI forecasting across multiple temporal scales, showing that the optimized LSTM outperformed an unoptimized LSTM and several baselines on standard error metrics [35]. Wei et al. proposed a WOA-CNN-LSTM-AM model that integrates whale optimization with a CNN-LSTM backbone and an attention mechanism for daily PM_{2.5} prediction, achieving a strong MAE/RMSE and goodness-of-fit [14].

Compared with these studies, the key contribution of our framework is not merely to further reduce point-prediction errors, but to improve decision-oriented reliability by explicitly modeling the predictive uncertainty while maintaining a high accuracy. A common feature of the PSO-LSTM, LSTM-SSA, and WOA-CNN-LSTM-AM is that they produce deterministic single-valued forecasts, and their evaluation is largely based on accuracy metrics such as the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Under abrupt pollution transitions, anomalous emissions, or increased observation noise, such point estimates cannot directly answer decision-critical questions such as whether a forecast is trustworthy or whether an early warning should be issued. In contrast, by Bayesianizing the LSTM, model weights are learned as probability distributions and sampled during inference, allowing parameter uncertainty to propagate into predictive uncertainty. Building on this, flipout introduces pseudo-independent perturbations for individual samples within each mini-batch, substantially reducing the gradient variance and alleviating training instability on large-scale time-series data, thereby enabling smoother updates and more consistent convergence. Moreover, the adopted standard-deviation calibration aligns the predictive intervals with the empirical error distribution, improving the statistical consistency and interpretability when confidence intervals are used for anomaly identification or threshold-based warning evaluations. Overall, relative to the deterministic, optimization-driven forecasting route represented by the above studies, our FBA-LSTM provides additional value by delivering a predictive accuracy alongside reliable uncertainty estimates for early warning and risk-sensitive applications.

4.2. Computational Cost and Convergence Characteristics

In the Bayesian training process, compared to traditional deterministic LSTM models, Bayesian LSTM typically requires longer training times. Bayesian inference methods require the precise posterior estimation of model parameters, which often leads to a significant increase in the computational costs. For example, as shown in Table 8, the training time for both Bayesian LSTM models was several hours, while other deterministic models required only a few minutes. However, the inference time difference was minimal, with at most a one-second difference.

Table 8. Computational cost and training-loss statistics of different models under 150 epochs.

Model	Training Time (min)	Inference Time (s)	Mean Training Loss	Training Loss Variance
LSTM	5	0.1	1.54×10^{-3}	1.87×10^{-3}
Residual LSTM	6	0.1	1.19×10^{-3}	2.67×10^{-3}
Attention LSTM	6	0.1	1.88×10^{-3}	2.61×10^{-3}
MC Dropout LSTM	7	0.2	2.79×10^{-3}	2.86×10^{-3}
Bayesian LSTM	254	1.1	6.85×10^4	5.39×10^4
FBA-LSTM	186	0.8	6.74×10^{-6}	1.12×10^{-6}

In the model design, we adopted the flipout technique to enhance the training efficiency of the Bayesian LSTM. The flipout method updates the model's weights using pseudo-independent perturbations during each forward pass, reducing the gradient vari-

ance and avoiding redundant forward passes inherent in traditional Bayesian methods. This approach accelerates the Bayesian inference process while improving the stability of training, resulting in smoother convergence and faster training times.

According to the statistics in Table 8, the flipout LSTM model achieved a significantly lower training loss compared to the traditional Bayesian LSTM model. Specifically, the mean training loss for the flipout LSTM model was 6.74×10^{-6} with a standard deviation of 1.12×10^{-6} , showing a steady decrease in loss with narrower fluctuations. In contrast, the traditional Bayesian LSTM model exhibited a much higher mean training loss of 6.85×10^4 with a larger standard deviation of 5.39×10^4 and a wider fluctuation range, resulting in slower convergence and making the training process harder to stabilize. These results demonstrate that the flipout method not only accelerates the Bayesian inference process, but also enhances the stability of model training, offering a more efficient and practical solution for complex Bayesian inference tasks.

4.3. Challenges and Future Directions in the Practical Application of FBA-LSTM

In $PM_{2.5}$ prediction, introducing uncertainty quantification can enhance the decision support effectiveness and model robustness. For example, Bashardoosta et al. demonstrated that incorporating uncertainty quantification (via a Bayesian neural network and Shannon entropy methods) in $PM_{2.5}$ spatial predictions helps identify high-reliability prediction regions and quantify spatial uncertainty, thereby providing targeted decision-making guidance for pollution mitigation in high-uncertainty urban border areas of Tehran [36]. Their study showed that the uncertainty-aware BNN-CNN model achieved a 95% accuracy in parameter-calibrated prediction, significantly outperforming the deterministic CNN model (82.7% accuracy) in noisy data scenarios, which confirms the value of uncertainty quantification in improving the decision-making reliability. Additionally, Zhou et al. showed that introducing uncertainty via variational inference can improve the model robustness and reduce prediction errors in multi-step $PM_{2.5}$ forecasting, especially when handling dynamic changes in meteorological and pollution factors [37]. By quantifying the predictive uncertainty through approximate Bayesian methods, their model generated reliable prediction intervals and reduced the risk of misleading predictions compared to traditional deterministic LSTM models. Both studies employed variational inference or Bayesian frameworks, which introduce uncertainty through latent variables or parameter posterior distributions. This proves that the fully Bayesian FBA-LSTM has a broad applicability, especially when handling sudden pollution events or dynamically changing environments, and shows a stronger robustness and accuracy compared to traditional deterministic models.

While the model achieved satisfactory results in the experiments, some limitations remain for real-world deployment. Although the introduction of the flipout technique helps mitigate the increase in the computational cost, further optimization is still required to enhance the scalability in practical applications. Future research should explore advanced techniques to improve the model scalability, such as model compression methods like network pruning or knowledge distillation [38], which can reduce the inference time without sacrificing the prediction accuracy. However, this introduces challenges in maintaining reliable uncertainty estimates during compression, which remains a key focus for future studies. Moreover, the current model uses data from a single monitoring station and lacks spatial information, which could limit the model's ability to detect and predict pollution dispersion in complex urban and industrial environments. Adding spatial features through multi-site joint modeling can enhance the model's performance by incorporating regional pollution patterns. Incorporating spatial awareness models, such as graph neural networks (GNNs), can improve the model's ability to handle multi-source interference and increase its sensitivity to pollution dynamics in neighboring regions [34].

In summary, the FBA-LSTM model is particularly suitable for urban air quality early warning systems where false alarms incur economic costs (e.g., unnecessary traffic restrictions) and missed warnings pose health risks. By incorporating Bayesian inference, attention mechanisms, and calibration strategies, the model demonstrates a strong predictive stability and early warning capability in complex pollution scenarios. However, addressing challenges related to missing spatial information, computational efficiency bottlenecks, and model deployment adaptability remains critical for enhancing its generalization ability and practical value in multi-source heterogeneous data environments.

5. Conclusions

In this paper, the FBA-LSTM model is proposed to predict the hourly PM_{2.5} concentrations at four monitoring stations in Nanjing for the year 2020, as well as to perform tasks related to pollution warning and anomaly detection. The experimental results demonstrate that the model generally outperformed comparative methods across several evaluation metrics. The key findings are summarized as follows:

- (a) Under both noisy and noise-free conditions, the FBA-LSTM model generally showed a better prediction performance across all four monitoring stations (1151A, 1157A, 1158A, and 1159A). Its RMSE and MAE values were lower than those of the traditional LSTM, attention LSTM, residual LSTM, and MC dropout LSTM, indicating a stronger accuracy and robustness.
- (b) In the pollution event warning task, the FBA-LSTM reduced both false alarms and omissions, maintaining a high accuracy even under boundary samples and high uncertainty. It demonstrated stronger uncertainty quantification and warning credibility compared to the other four models.
- (c) In the task of detecting anomalous pollution events, the FBA-LSTM outperformed traditional detection methods such as isolation forest (IF) and the interquartile range (IQR). It achieved optimal results in terms of the accuracy, recall, and F1 scores, indicating its potential for recognizing emerging pollution anomalies.

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

Funding: This research was funded by the National Natural Science Foundation of China, grant number 41478324.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

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

References

1. Zhang, X.; Chen, X.; Yue, Y.; Wang, S.; Zhao, B.; Huang, X.; Li, T.; Sun, Q.; Wang, J. Ecological study on global health effects due to source-specific ambient fine particulate matter exposure. *Environ. Sci. Technol.* **2023**, *57*, 1278–1291. [[CrossRef](#)]
2. Qing, W. Urbanization and global health: The role of air pollution. *Iran. J. Public Health* **2018**, *47*, 1644.
3. Zhang, Y.; Tian, J.; Wang, Q.; Qi, L.; Manousakas, M.I.; Han, Y.; Ran, W.; Sun, Y.; Liu, H.; Zhang, R.; et al. High-time-resolution chemical composition and source apportionment of PM 2.5 in northern Chinese cities: Implications for policy. *Atmos. Chem. Phys.* **2023**, *23*, 9455–9471. [[CrossRef](#)]

4. Gao, M.; Saide, P.E.; Xin, J.; Wang, Y.; Liu, Z.; Wang, Y.; Wang, Z.; Pagowski, M.; Guttikunda, S.K.; Carmichael, G.R. Estimates of health impacts and radiative forcing in winter haze in eastern China through constraints of surface PM_{2.5} predictions. *Environ. Sci. Technol.* **2017**, *51*, 2178–2185. [[CrossRef](#)]
5. Lv, B.; Cobourn, W.G.; Bai, Y. Development of nonlinear empirical models to forecast daily PM_{2.5} and ozone levels in three large Chinese cities. *Atmos. Environ.* **2016**, *147*, 209–223. [[CrossRef](#)]
6. Semlali, B.E.B.; El Amrani, C.; Ortiz, G.; Boubeta-Puig, J.; Garcia-de Prado, A. SAT-CEP-monitor: An air quality monitoring software architecture combining complex event processing with satellite remote sensing. *Comput. Electr. Eng.* **2021**, *93*, 107257. [[CrossRef](#)]
7. Semlali, B.E.B.; Amrani, C.E. A stream processing software for air quality satellite datasets. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2020)*, *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development, Agadir, Morocco, 17–23 December 2020*; Springer: Cham, Switzerland, 2020; pp. 839–853.
8. Liu, D.j.; Li, L. Application study of comprehensive forecasting model based on entropy weighting method on trend of PM_{2.5} concentration in Guangzhou, China. *Int. J. Environ. Res. Public Health* **2015**, *12*, 7085–7099. [[CrossRef](#)]
9. Ausati, S.; Amanollahi, J. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM_{2.5}. *Atmos. Environ.* **2016**, *142*, 465–474. [[CrossRef](#)]
10. Huang, K.; Xiao, Q.; Meng, X.; Geng, G.; Wang, Y.; Lyapustin, A.; Gu, D.; Liu, Y. Predicting monthly high-resolution PM_{2.5} concentrations with random forest model in the North China Plain. *Environ. Pollut.* **2018**, *242*, 675–683. [[CrossRef](#)] [[PubMed](#)]
11. Liu, W.; Guo, G.; Chen, F.; Chen, Y. Meteorological pattern analysis assisted daily PM_{2.5} grades prediction using SVM optimized by PSO algorithm. *Atmos. Pollut. Res.* **2019**, *10*, 1482–1491. [[CrossRef](#)]
12. Bedi, S.; Katiyar, A.; Krishnan, N.A.; Kota, S.H. Utilizing LSTM models to predict PM_{2.5} levels during critical episodes in Delhi, the world's most polluted capital city. *Urban Clim.* **2024**, *53*, 101835. [[CrossRef](#)]
13. Zrira, N.; Kamal-Idrissi, A.; Farssi, R.; Khan, H.A. Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism. *J. Sea Res.* **2024**, *198*, 102472. [[CrossRef](#)]
14. Wei, Q.; Zhang, H.; Yang, J.; Niu, B.; Xu, Z. PM_{2.5} concentration prediction using a whale optimization algorithm based hybrid deep learning model in Beijing, China. *Environ. Pollut.* **2025**, *371*, 125953. [[CrossRef](#)]
15. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589. [[CrossRef](#)]
16. Sadr, M.A.M.; Zhu, Y.; Hu, P. An anomaly detection method for satellites using Monte Carlo dropout. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *59*, 2044–2052. [[CrossRef](#)]
17. Tabas, S.S.; Samadi, S. Variational Bayesian dropout with a Gaussian prior for recurrent neural networks application in rainfall–runoff modeling. *Environ. Res. Lett.* **2022**, *17*, 065012. [[CrossRef](#)]
18. Han, Y.; Lam, J.C.; Li, V.O.; Reiner, D. A Bayesian LSTM model to evaluate the effects of air pollution control regulations in Beijing, China. *Environ. Sci. Policy* **2021**, *115*, 26–34. [[CrossRef](#)]
19. Hubin, A.; Storvik, G. Sparse Bayesian neural networks: Bridging model and parameter uncertainty through scalable variational inference. *Mathematics* **2024**, *12*, 788. [[CrossRef](#)]
20. Chandra, R.; Simmons, J. Bayesian neural networks via MCMC: A Python-based tutorial. *IEEE Access* **2024**, *12*, 70519–70549. [[CrossRef](#)]
21. Wen, Y.; Vicol, P.; Ba, J.; Tran, D.; Grosse, R. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv* **2018**, arXiv:1803.04386. [[CrossRef](#)]
22. Jouppi, N.P.; Young, C.; Patil, N.; Patterson, D.; Agrawal, G.; Bajwa, R.; Bates, S.; Bhatia, S.; Boden, N.; Borchers, A.; et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, Toronto, ON, Canada, 24–28 June 2017; pp. 1–12.
23. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
24. Yu, Y.; He, S.; Wu, X.; Zhang, C.; Yao, Y.; Liao, H.; Wang, Q.; Xie, M. PM_{2.5} elements at an urban site in Yangtze River Delta, China: High time-resolved measurement and the application in source apportionment. *Environ. Pollut.* **2019**, *253*, 1089–1099. [[CrossRef](#)] [[PubMed](#)]
25. Huang, L.; Zhou, L.; Chen, J.; Chen, K.; Liu, Y.; Chen, X.; Tang, F. Acute effects of air pollution on influenza-like illness in Nanjing, China: A population-based study. *Chemosphere* **2016**, *147*, 180–187. [[CrossRef](#)]
26. Yang, H.; Ge, A.; Xie, H.; Li, W.; Qin, Y.; Yang, W.; Wang, D.; Gu, W.; Wang, X. Effects of ambient air pollution on precocious puberty: A case-crossover analysis in Nanjing, China. *J. Clin. Med.* **2022**, *12*, 282. [[CrossRef](#)]
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
28. Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
29. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the International Conference on Machine Learning*, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2796–2804.

30. Cui, P.; Hu, W.; Zhu, J. Calibrated reliable regression using maximum mean discrepancy. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 17164–17175.
31. Chung, Y.; Char, I.; Guo, H.; Schneider, J.; Neiswanger, W. Uncertainty Toolbox: An Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification. *arXiv* **2021**, arXiv:2109.10254. [[CrossRef](#)]
32. Aksangür, İ.; Eren, B.; Erden, C. Evaluation of data preprocessing and feature selection process for prediction of hourly PM₁₀ concentration using long short-term memory models. *Environ. Pollut.* **2022**, *311*, 119973. [[CrossRef](#)]
33. HJ 633-2012; Technical Regulation on Ambient Air Quality Index. Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2012.
34. Chen, Y.; Wu, Y.; Zhang, S.; Yuan, K.; Huang, J.; Shi, D.; Hu, S. Regional PM_{2.5} prediction with hybrid directed graph neural networks and Spatio-temporal fusion of meteorological factors. *Environ. Pollut.* **2025**, *366*, 125404. [[CrossRef](#)]
35. Liu, X.; Guo, H. Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai. *Atmos. Pollut. Res.* **2022**, *13*, 101551. [[CrossRef](#)]
36. Bashardoost, A.; Mesgari, M.S.; Karimi, M. Quantifying uncertainty in the spatial prediction of PM_{2.5} using a deep learning algorithm. *Ann. GIS* **2025**, *31*, 679–712. [[CrossRef](#)]
37. Murad, A.; Kraemer, F.A.; Bach, K.; Taylor, G. Probabilistic deep learning to quantify uncertainty in air quality forecasting. *Sensors* **2021**, *21*, 8009. [[CrossRef](#)] [[PubMed](#)]
38. Magris, M.; Iosifidis, A. Bayesian learning for neural networks: An algorithmic survey. *Artif. Intell. Rev.* **2023**, *56*, 11773–11823. [[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.