DOI: 10.37943/24WXPA8545

Saltanat Sharipova
Manager of the Center of competence and excellence
saltanat.sharipova@astanait.edu.kz, orcid.org/0000-0001-7267-3261
Astana IT University, Kazakhstan

Dilara Abzhanova
Director of the Center of competence and excellence
dilara.abzhanova@astanait.edu.kz, orcid.org/0000-0002-7988-3971
Astana IT University, Kazakhstan

Sapar Toxanov
Vice-rector for Educational work
sapar.toxanov@astanait.edu.kz, orcid.org/0000-0002-2915-9619
Astana IT University, Kazakhstan

Andrii Biloshchytskyi
Vice-rector for Science and Innovation
ab@astanait.edu.kz, orcid.org/0000-0001-9548-1959
Astana IT University, Kazakhstan

DEVELOPMENT OF A NEURAL NETWORK-BASED MODULE FOR FORECASTING ATMOSPHERIC POLLUTANT EMISSIONS

Abstract: The prediction of emissions to air is a crucial and complex task for environmental monitoring and air quality management. Accurate forecasting is essential for the timely adoption of mitigation measures and for ensuring regulatory compliance. However, traditional statistical methods often perform inadequately because they poorly capture non-linear dependencies, intricate interactions between variables, and long-term temporal patterns, all of which ultimately decrease forecasting accuracy.

The work presents an emission prediction software module based on a neural network with LSTM architecture. The input factors used were the concentrations of the main pollutants (NO, NO_2, SO_2, CO, solid particles) as well as meteorological indicators including air temperature, humidity and flow rate. Data provided by the operating enterprises, including 39,803 lines with increments of 20 minutes, were pre-processed: cleared from skips, normalized parameters and forming training sequences of 72 steps, corresponding to the daily interval. Additional exploration analysis was performed, which revealed the presence of expressed daily and weekly cycles, as well as correlations between weather conditions and concentrations of pollutants.

The built model showed the ability to reproduce emission dynamics with acceptable accuracy, which is confirmed by MSE 0.87 and R^2 0.86 values. The developed module is integrated into the current monitoring system and provides a user-friendly interface for building real-time forecasts. The results are consistent with current research, but the work is applied as a tool used in practical activities. In the future, it is planned to expand the set of factors and explore the possibilities of using ensemble architecture to improve the accuracy and robustness of forecasts.

Keywords: neural networks; air pollution forecasting; atmospheric pollutant emissions; environmental monitoring; predictive modeling.

Received: 27.08.2025 Accepted: 20.10.2025 Published: 30.10.2025

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

Introduction

Air pollution is one of the current environmental problems affecting public health, ecosystems and climate resilience. Pollutant emission forecasting plays an important role in environmental monitoring and supporting management decisions related to the transition to sustainable development and carbon neutrality.

Traditional statistical methods, such as the auto-regression model of the moving average (ARIMA), and machine learning algorithms, including random forest and gradient boosting, are used in atmospheric emission forecasting [1]. However, these methods struggle to model nonlinear temporal dependencies, especially when both short- and long-term patterns must be considered [2], [3].

In this context, models of deep learning are attracting more attention, in particular the LSTM (Long Short-Term Memory) architecture. These models are capable of processing sequential data and capturing time dependencies at different scales, making them suitable for the analysis of air pollution dynamics.

For example, in the study [4], the RNN-LSTM model developed to predict PM concentrations. showed low error values and confirmed the applicability of neural network approaches with limited data. The paper [5] states that neural network and ensemble methods incorporating LSTM provide higher accuracy than traditional algorithms.

The comparison of LSTM with classical statistical models also shows its advantages. Scientists [6] have found that the application of LSTM to predict PM concentrations. Dakar allowed the RMSE to be reduced almost by half compared with the ARIMA model. Similar results were obtained in a paper [7], where LSTM provided higher prediction quality than the multi-layered neural network (ANN) when modelling ozone concentrations.

A separate direction is related to the creation of hybrid solutions. Thus, a separate direction is related to the creation of hybrid solutions that combine space-time structure and attention. In this context, the STA-ResConvLSTM (Stationary-Temporal Attention ResNet-ConvLSTM) model showed improved prediction accuracy in urban agglomerations, including areas around Beijing, surpassing basic models on RMSE, MAE and R^2 [8]. In addition, other scientists [9] have proposed the VMD-GAT-BiLSTM hybrid architecture, which has improved the prediction of pollutant concentrations.

Systematic reviews of recent years confirm that deep neural networks and their ensemble variants show better results than traditional methods. For example, work [10] proposed two-channel deep learning model combines visual and numerical data for $PM_{2.5}$ and PM_{10} prediction with R^2 0.946/0.905 and RMSE 4.79/11.51 μg/m3 respectively, which confirms the prospect of multi-source approaches, and the study [11] points growing interest in hybrid and transformative architectures for air pollution monitoring and forecasting.

Special attention is given to transformative architecture. Scientists [12] proposed the AirFormer model for predicting air quality across a country. Similar conclusions are presented in a paper [13], where the PlumeNet model based on ConvLSTM has shown efficiency in large-scale forecasting. Researchers [14] have shown the effectiveness of CNN-LSTM for predicting PM concentrations in urban agglomerations. The review [15] of machine learning methods also identifies LSTM and hybrid architecture as the most promising for environmental forecasting. The paper [16] confirmed the effectiveness of LSTM in short-term air pollution forecasting in South-East Asia.

Therefore, LSTM architectures and their hybrid variants (CNN-LSTM, GAT-LSTM, Transformer-based models) demonstrate strong potential for atmospheric emission forecasting. Using this approach, atmospheric emissions can be predicted with high accuracy.

Although many studies confirm the effectiveness of deep learning for air pollution forecasting, most existing research still focuses primarily on algorithmic performance evaluation

DOI: 10.37943/24WXPA8545

using historical datasets only [4], [5], [6], [7], [8]. Limitations include insufficient analysis of real industrial environments, limited temporal resolution, narrow feature engineering, and the absence of deployment-oriented validation. Moreover, many studies do not integrate forecast modules into operational environmental monitoring platforms, which restricts the applicability for real-time decision-making.

There is a lack of practically implemented solutions that forecast emissions based on real-time industrial data streams while accounting for combined meteorological and operational factors.

Based on this gap, the present work proposes and evaluates a fully integrated forecasting module operating within an environmental monitoring ecosystem in Kazakhstan.

Problem statement: Current forecasting approaches are insufficiently adapted for nonlinear industrial emission dynamics and often lack real-time applicability for operational environmental monitoring.

Research aim: To develop and validate a neural-network-based forecasting module for pollutant emissions integrated into an operational monitoring system.

Objectives:

(1) Conduct data analysis to identify dominant temporal and meteorological factors;

(2) Design and train an LSTM-based forecasting model;

(3) Compare predictive performance with classical baselines;

(4) Deploy and validate the module within an existing monitoring system.

Hypothesis: Deep learning methods (LSTM) can significantly improve short-term forecasting accuracy compared to traditional statistical approaches (e.g., ARIMA) when applied to industrial monitoring data.

Methods and Materials

For the development of the forecasting module, data provided by active enterprises in Kazakhstan were used. The dataset included 39,803 time-stamped observations recorded at 20-minute intervals. Each observation contained both pollutant concentrations and relevant meteorological indicators. Primary pollutants (nitrogen oxides, sulphur dioxide, carbon monoxide, particulate matter), meteorological indicators (air temperature, humidity, wind speed and direction) and additional temporal indicators were considered as inputs (hours, days of the week, daily cycles). This set of factors allowed for modelling the dependence of emissions on weather conditions and time patterns [17].

Before building the model, pre-processing of data was performed. At this stage records with missing or incorrect values were deleted, time stamps were given a single interval of 20 minutes, and numerical parameters were normalized in the range [0.1] by MinMax-scaling. Training sequences of 72 steps (24 hours) were formed for the training, which allowed to take into account daily fluctuations of emissions and use them in the forecast.

In addition, an exploratory data analysis (EDA) was carried out to identify patterns and characteristics of the sample structure. Time series graphs, distribution histograms and correlation matrices were developed as part of the analysis. The results showed marked daily and weekly seasonality, dependence of concentration on weather conditions and episodic peaks indicating irregular spikes in emissions. These observations are taken into account in the choice of model architecture and learning strategy.

The neural network architecture LSTM (Long Short-Term Memory) was used for the prediction, which is distinguished by its ability to model time dependencies of different durations. Unlike traditional recursive networks (RNN), LSTM effectively addresses the problem of disappearing and exploding gradients thanks to a memory mechanism [18], [19].

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

The basic element of LSTM consists of a memory cell and three gates: forgetting, entry and exit. Their work is formalized by the following equations (formula 1):

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \\ h_t &= o_t \cdot \tanh(C_t), \end{aligned} \tag{1}$$

where x_t – input data at time t, h_t – hidden state, C_t – memory cell state, f_t, i_t, o_t – value of the gates, σ – sigmoidal activation function, $tanh$ – hyperbolic activation function.

The architecture of the built model (Figure 1) included an input layer, the first LSTM layer with 32 neurons for long-term dependency extraction, a dropout layer with a 0.2 factor for regularization, The second LSTM layer on 16 neurons for short-term patterns and the final fully connected Dense layer with one neuron predicting the next time step. This structure allowed to take into account both long-term and short-term fluctuations of emissions, maintaining the model's resistance to re-training.

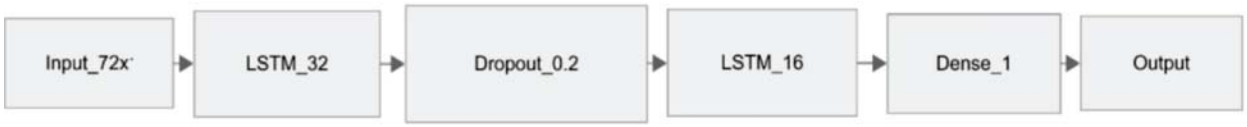

Figure 1. Neural network architecture.

The standard time series analysis metrics used in time series studies [20]: standard error (MSE) (formula 2) and coefficient of determinism (R^2) (formula 3) were applied to quantify the quality of the forecasts. They were calculated by the following formulas:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{3}$$

where y_i – actual values, $\hat{y}_i$ – forecast values, $\bar{y}$ – sample average.

To illustrate the structure of the network, an architectural diagram was prepared, including all the main layers and connections (Fig. 1).

For reproducibility, a classical ARIMA baseline was trained using the same dataset and AIC model selection strategy. Statistical significance of improvements was evaluated using a paired t-test ($\alpha = 0.05$).

Results

Exploration data analysis (EDA) showed the presence of expressed time patterns. The time series clearly show daily and weekly cycles, as well as individual concentration peaks that may be associated with man-made emissions or changes in weather conditions. Correlation analysis revealed significant relationships between pollutant levels and meteorological parameters, in particular temperature, humidity and airflow rate.

DOI: 10.37943/24WXPA8545

Table 1. Descriptive sample statistics (n = 39 803 lines)

Parameter	Average	Minimum	Maximum	St. deviation
NO	0.197	0.150	0.296	0.026
NO_2	0.310	0.250	0.424	0.033
SO_2	0.118	0.090	0.150	0.014
CO	1.695	1.500	1.986	0.100
DUST	39.4	36.1	43.2	1.1
O_2	20.8	20.6	21.1	0.1
TEMP (°C)	9.0	-18.5	32.1	9.5
Humidity (%)	63.2	15.0	96.2	18.3

Figure 2 shows a fragment of the time series of NO_2 concentrations for the first 500 observations. Values appear to fluctuate within normalized ranges and reflect daily dynamics.

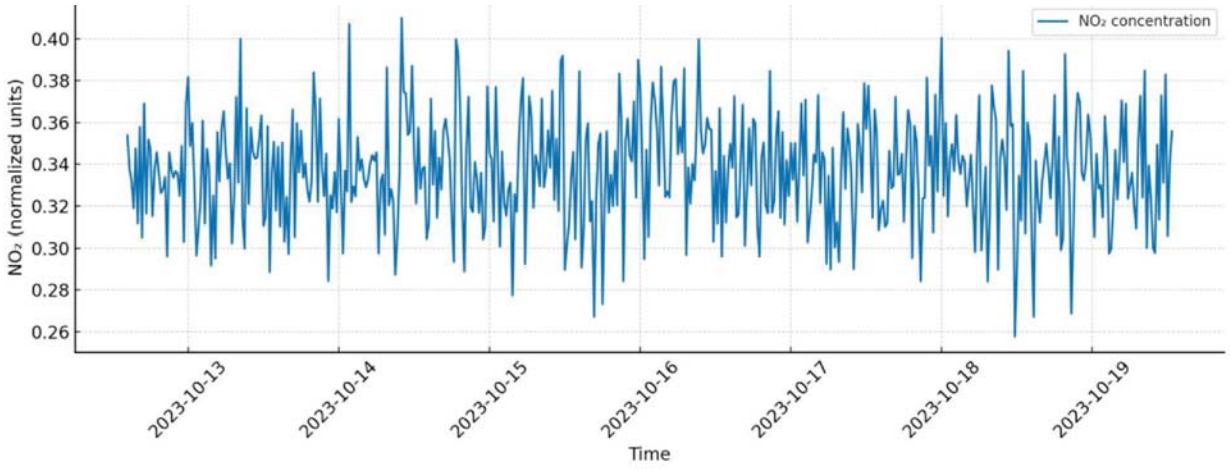

Figure 2. Time range of concentrations NO_2 (the first 500 observations).

The distribution of pollutants is characterized by asymmetry and peak values. For example, the dust histogram (DUST) shown in Figure 3 indicates a shift of the distribution to high values, which may be related to local emissions.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

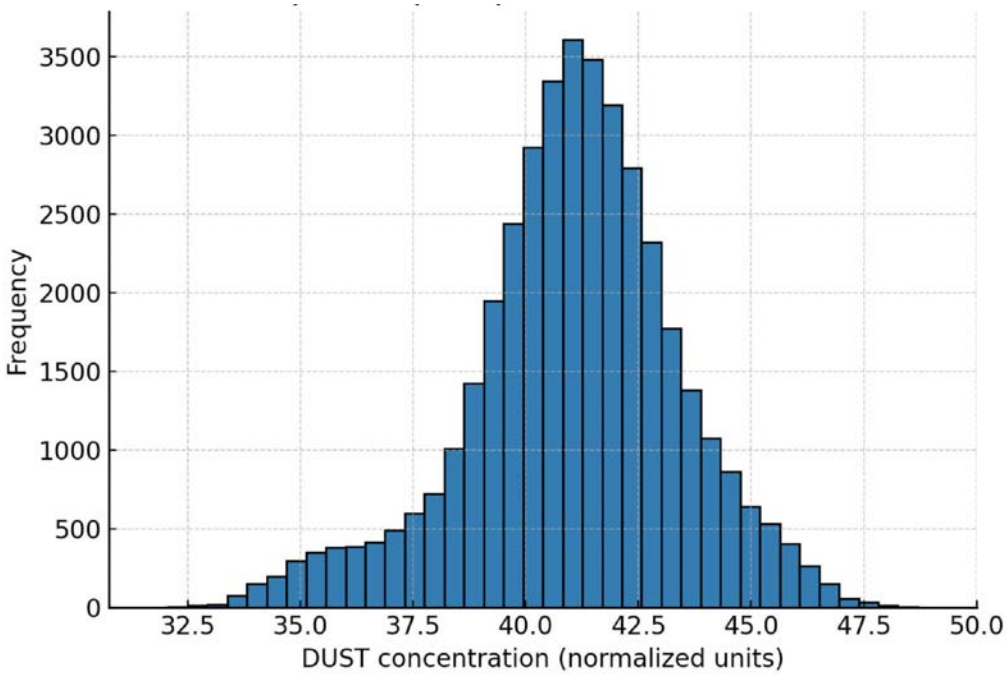

Figure 3. Dust distribution histogram (DUST).

For a more detailed evaluation of the relationships between factors, a correlation matrix was constructed (figure 4). The highest correlation factors are found between NO and NO_2 gas concentrations, as well as between humidity and dust. This confirms the need to take into account an integrated set of input factors in forecasting.

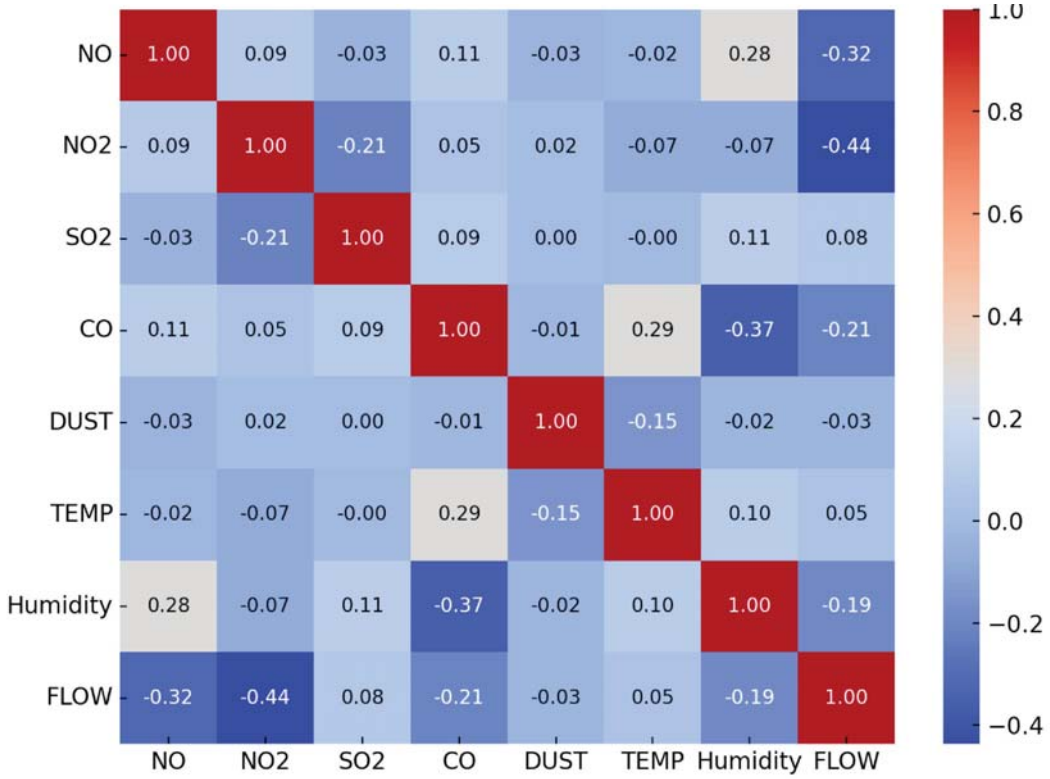

Figure 4. Correlation matrix of pollution factors and meteorological parameters.

DOI: 10.37943/24WXPA8545

The quality of the developed model was assessed using standard error (MSE) and coefficient of determination (R^2). The final values were MSE 0.87 and R^2 0.86, which shows the model's ability to reproduce the behavior of time series and make predictions based on historical data.

Table 2 summarizes the predictive performance of the proposed LSTM model relative to the classical ARIMA baseline. The LSTM achieved a 32–48% reduction in RMSE depending on the pollutant, and the improvement was statistically significant (paired t-test, $p < 0.05$).

Table 2. Comparison of proposed model with ARIMA

Pollutant	ARIMA RMSE	LSTM RMSE	Improvement (%)	p-value
NO	0.145	0.088	39.3%	< 0.05
NO_2	0.158	0.084	46.8%	< 0.05
SO_2	0.121	0.081	33.1%	< 0.05
CO	0.210	0.131	37.6%	< 0.05
DUST	0.274	0.173	36.9%	< 0.05

The LSTM approach consistently outperformed ARIMA across all pollutants with RMSE reduction ranging from 33% to 47%, indicating a substantially improved ability to model nonlinear temporal emission dynamics. The highest relative improvement was observed for NO_2, which exhibits pronounced short-term variability driven by industrial processes and meteorological fluctuations.

This confirms that deep learning-based forecasting provides a stronger capability to model nonlinear emission dynamics compared to traditional statistical methods.

Figure 5 shows the interface of the forecast module, allowing the user to select the required indicator and forecast horizon.

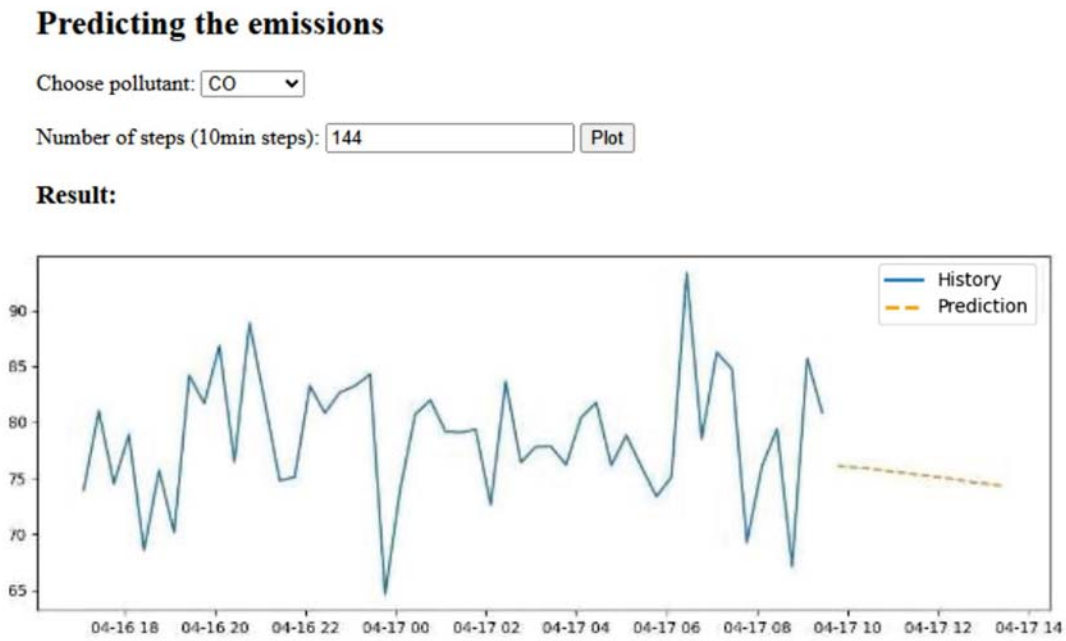

Figure 5. Prediction module interface.

Automatically creates a graph where the forecast values are superimposed on the latest actual observations, which makes it easier to understand the results.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

In addition, the stability of the model on individual sub-samples was tested. The metric values obtained were kept at a comparable level, which indicates that the module is stable. It is important to note that the model correctly reproduces both seasonal fluctuations and abrupt emissions of concentrations, which confirms its applicability for practical environmental monitoring tasks.

Discussion

The presented emission forecasting module has been successfully implemented into the existing environmental monitoring system and demonstrated the ability to produce forecasts with acceptable accuracy. An important contribution of this work is not only the forecasting accuracy, but also the practical implementation of the tool for real-time use. The module interface allows specialists to quickly analyze the current situation and make predictions for different horizons, which increases the value of the solution for air quality management applications.

Comparison with existing studies shows that most of the work is limited to theoretical modeling and testing of algorithms on historical data. For example, a study by Gueye and co-authors (2025) examined the effectiveness of the LSTM model for predicting PM concentrations. in Dakar, and the authors showed a reduction of errors compared to classical statistical methods [3]. However, their work was focused on evaluating the accuracy of the algorithm and did not include integrating the solution into real monitoring systems. In contrast, the module developed by us combines the experimental part with the application implementation, which allows to use the simulation results directly in management processes.

It should be noted that in the current version of the module only the main pollutants and basic meteorological parameters are taken into account as input factors. A wider consideration of external conditions, such as seasonal changes in traffic flows or features of industrial facilities, can further improve the accuracy of forecasts and broaden the scope of application of the module.

The model currently operates at a single monitoring location with a limited set of meteorological variables. External influences such as industrial operational modes, traffic intensity, and seasonal technological cycles are not yet explicitly modeled. These aspects can contribute to significant emission fluctuations and therefore remain an important direction for future improvement.

The methodology can be extended toward multi-site learning across different industrial zones, as well as incorporating additional environmental and operational datasets. Further enhancement may be achieved through ensemble architecture and adaptive models capable of handling spatio-temporal variability. These improvements will increase robustness and enable nationwide deployment of real-time emission forecasting tools.

Thus, the key contribution of this study is not only the construction of a prognostic model, but also the creation of a software component that has been integrated into the current intellectual system. This confirms the practical relevance of the proposed solution and opens opportunities for further development, such as multi-factor forecasting and ensemble architecture.

Conclusion

The work presents a pollutant emission prediction module, developed on the basis of a neural network with LSTM architecture and integrated into an existing environmental monitoring system. The analysis showed that the model provides an acceptable accuracy of predictions (MSE 0.87, R^2 0.86) and is able to reproduce both smooth seasonal fluctuations and sudden spikes in pollutant concentrations.

An important result is not only the successful construction of the model, but also its implementation in the form of a software module with a convenient interface. This allows professionals to use real-time forecasts in environmental management decisions.

Each research objective was successfully addressed:

- The EDA confirmed the presence of temporal and meteorological dependencies affecting emission dynamics;
- The trained LSTM model demonstrated high predictive capability (MSE 0.87, R^2 0.86);
- When compared with the ARIMA baseline, the proposed model showed statistically significant improvement in forecasting accuracy;
- Integration into the monitoring software validated the practical applicability for real-time environmental decision-support.

These results confirm that the formulated research gap has been successfully closed and the developed module represents a valuable contribution to operational air quality forecasting systems.

Acknowledgment

This paper was written within the framework of the state order for the implementation of a scientific program in accordance with the budget program 217 “Development of Science”, IRN No. BR21882258 on the topic: "Development of a complex of intelligent information and communication systems for environmental monitoring of emissions into the environment for making management decisions in the concept of carbon neutrality."

References

[1] Guo, Q., He, Z., Li, S., Li, X., Meng, J., Hou, Z., ... & Chen, Y. (2020). Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol and Air Quality Research*, *20*(6), 1429-1439.

[2] Petrić, V., Hussain, H., Časni, K., Vuckovic, M., Schopper, A., Andrijić, Ž. U., ... & Lovrić, M. (2024). Ensemble machine learning, deep learning, and time series forecasting: improving prediction accuracy for hourly concentrations of ambient air pollutants. *Aerosol and Air Quality Research*, *24*(12), 230317.

[3] Chen, M., Xu, P., Liu, Z., Liu, F., Zhang, H., & Miao, S. (2025). Air pollution prediction based on optimized deep learning neural networks: PSO-LSTM. *Atmospheric Pollution Research*, *16*(3), 102413.

[4] Bhalgat, P., Pitale, S., & Bhoite, S. (2019). Air quality prediction using machine learning algorithms. International Journal of Computer Applications Technology and Research, 8(9), 367-370.

[5] He, Z., & Guo, Q. (2024). Comparative analysis of multiple deep learning models for forecasting monthly ambient PM2. 5 concentrations: A case study in Dezhou City, China. *Atmosphere*, *15*(12), 1432.

[6] Gueye, A., Niang, S. A. A., Diallo, I., Dramé, M. S., Diallo, M., & Younous, A. A. (2025). On the Application of Long Short-Term Memory Neural Network for Daily Forecasting of PM2. 5 in Dakar, Senegal (West Africa). *Sustainability*, *17*(12), 5421.

[7] Guo, Q., He, Z., & Wang, Z. (2025). Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in Liaocheng City. *Scientific Reports*, *15*(1), 6798.

[8] Chen, C., Qiu, A., Chen, H., Chen, Y., Liu, X., & Li, D. (2023). Prediction of pollutant concentration based on spatial–temporal attention, ResNet and ConvLSTM. *Sensors*, *23*(21), 8863.

[9] Wang, X., Zhang, S., Chen, Y., He, L., Ren, Y., Zhang, Z., ... & Zhang, S. (2024). Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM. *Scientific Reports*, *14*(1), 17841.

[10] Wu, Y. (2024). Time-Series Forecasting of $PM_{2.5}$ and PM_{10} Concenntrations Using a Dual-Channel Deep Learning Model Integrating Visual and Numerical Data. *Sensors*, 25(1), 95.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 24, OCTOBER 2025

[11] Chadalavada, S., Faust, O., Salvi, M., Seoni, S., Raj, N., Raghavendra, U., ... & Acharya, R. (2025). Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review. *Environmental Modelling & Software*, *185*, 106312.
[12] Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., ... & Zimmermann, R. (2023, June). Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 12, pp. 14329-14337).
[13] Alléon, A., Jauvion, G., Quennehen, B., & Lissmyr, D. (2020). PlumeNet: Large-scale air quality forecasting using a convolutional LSTM network. *arXiv preprint arXiv:2006.09204*.
[14] Bai, X., Zhang, N., Cao, X., & Chen, W. (2024). Prediction of PM2. 5 concentration based on a CNN-LSTM neural network algorithm. *PeerJ*, *12*, e17811.
[15] Özüpak, Y., Alpsalaz, F., & Aslan, E. (2025). Air Quality Forecasting Using Machine Learning: Comparative Analysis and Ensemble Strategies for Enhanced Prediction. *Water, Air, & Soil Pollution*, *236*(7), 464.
[16] Kristiani, E., Lin, H., Lin, J. R., Chuang, Y. H., Huang, C. Y., & Yang, C. T. (2022). Short-term prediction of PM2. 5 using LSTM deep learning methods. *Sustainability*, *14*(4), 2068.
[17] Zhang, Q., Han, Y., Li, V. O., & Lam, J. C. (2022). Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution estimation and forecast in metropolitan cities. *IEEE access*, *10*, 55818-55841.
[18] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[19] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
[20] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR).