AI-Powered Multi-City Air Quality Forecasting Platform

Venkata Tharun Rajana, Prof.(Dr.) Ravi Kiran, Rayudu Sadhana, Pandrangi Anjan Sai, Kambham Sunandh
Department of Computer Science & Engineering, Raghu Engineering College, Dakamarri(V), Visakhapatnam, Andhra Pradesh, India.

Abstract

Air pollution, particularly fine particulate matter (PM2.5), poses a significant public health threat to rapidly industrialising coastal cities. Visakhapatnam (Vizag), Andhra Pradesh, India, hosts major steel, petroleum, and port industries that contribute substantially to ambient PM2.5 concentrations. Accurate short-term forecasting of PM2.5 is essential for timely government advisories and public health interventions. This paper presents a comparative study of four predictive frameworks applied to hourly PM2.5 forecasting in Visakhapatnam: (i) Long Short-Term Memory (LSTM) networks, (ii) Transformer-based sequence models with multi-head self-attention, (iii) Extreme Gradient Boosting (XGBoost), and (iv) a weighted ensemble of all three. The models are trained on a two-year historical dataset comprising 17,520 hourly records of air quality and meteorological variables sourced from the Open-Meteo API. Forty-two engineered features — including temporal encodings, lag variables, rolling statistics, and meteorological interaction terms — are used as inputs to predict the next-hour PM2.5 concentration. Experimental results demonstrate that LSTM achieves the best individual performance (RMSE = 3.52 µg/m³, R^2 = 0.966), significantly outperforming Transformer (RMSE = 6.10, R^2 = 0.897) and XGBoost (RMSE = 6.61, R^2 = 0.879). The ensemble model attains RMSE = 4.45 µg/m³ and R^2 = 0.945, showing improved robustness over XGBoost and Transformer while remaining competitive with LSTM. The proposed system is integrated into a real-time Streamlit-based dashboard powered by a Groq LLM API for natural language air quality advisory generation. Results validate the superiority of deep sequential models for urban air quality time-series prediction and highlight the promise of hybrid ensemble approaches for robust operational deployment.

Keywords: PM2.5 forecasting, air quality prediction, LSTM, Transformer, XGBoost, ensemble learning, deep learning, time-series, Visakhapatnam, urban air pollution

1. Introduction

Particulate matter with an aerodynamic diameter of less than 2.5 micrometres (PM2.5) is recognised by the World Health Organisation (WHO) as one of the most hazardous atmospheric pollutants, associated with respiratory disease, cardiovascular disorders, and premature mortality [1]. Rapid urbanisation and industrial expansion in Indian coastal cities have exacerbated PM2.5 levels beyond safe limits, with Visakhapatnam frequently recording concentrations that exceed the National Ambient Air Quality Standards (NAAQS) of 60 µg/m³ for 24-hour averages [2]. Accurate short-term PM2.5 forecasting enables government bodies, municipal corporations, and public health agencies to issue timely warnings and trigger pollution-control protocols [3].

Traditional numerical air quality models, such as the Weather Research and Forecasting Model coupled with Chemistry (WRF-Chem) and AERMOD, rely on complex physical and chemical parameterisations that demand significant computational resources and detailed emission inventories [4]. In contrast, data-driven machine learning approaches have gained traction as practical alternatives, capable of learning non-linear pollutant dynamics directly from historical observations [5]. Among these, deep learning models — particularly recurrent architectures such as LSTM and attention-based Transformers — have demonstrated remarkable accuracy in modelling temporal dependencies inherent in air quality time-series data [6, 7].

Visakhapatnam, a port city on the eastern coast of India, presents a unique and challenging forecasting environment. Its industrial corridor encompasses steel manufacturing (Rashtriya Ispat Nigam Limited), a petroleum refinery (HPCL), shipbuilding yards, and a busy commercial port, all of which contribute diverse

and time-varying emission profiles. Sea-breeze dynamics, seasonal monsoon patterns, and orographic effects from the Eastern Ghats further modulate pollutant dispersion, making this a rich but complex case study [8]. Despite its industrial significance, very few machine learning studies have specifically targeted real-time PM2.5 forecasting for Visakhapatnam using multi-model comparative frameworks.

This paper addresses this gap by developing and benchmarking four distinct predictive frameworks using two years of hourly PM2.5 and meteorological data retrieved from the Open-Meteo historical API. The study makes the following primary contributions:

- A systematic comparison of LSTM, Transformer, XGBoost, and ensemble models for hourly PM2.5 forecasting under identical training and evaluation conditions.
- A comprehensive feature engineering pipeline producing 42 temporal, lag, rolling, and interaction features from raw multivariate time-series data.
- An ensemble strategy that combines LSTM, Transformer, and XGBoost predictions through weighted averaging, improving prediction stability across varying pollution events.
- A full-stack deployment framework integrating the prediction pipeline with a Streamlit dashboard and Groq LLM for natural language air quality advisory generation.

1.1 Literature Survey

1.1.1 Related Works

Air quality forecasting using machine learning has been an active area of research over the past decade. Early works employed shallow models such as Support Vector Regression (SVR) and Artificial Neural Networks (ANN) to capture non-linear relationships between meteorological variables and pollutant concentrations [9]. However, these approaches failed to adequately model the long-range temporal dependencies that characterise pollutant accumulation and dispersion events, motivating the adoption of recurrent deep learning architectures.

LSTM networks, introduced by Hochreiter and Schmidhuber [10], have become the de facto standard for sequential air quality modelling. Studies by Li et al. [11] demonstrated that LSTM-based models outperform traditional autoregressive models and feed-forward networks for PM2.5 prediction in Beijing, achieving R^2 values exceeding 0.93 over 24-hour forecasting horizons. Subsequent works extended LSTM with bidirectional processing and convolutional front-ends to enhance local pattern extraction [12].

The emergence of the Transformer architecture, originally proposed by Vaswani et al. [13] for natural language processing, has catalysed exploration of attention-based models for temporal forecasting. Temporal Fusion Transformers (TFT) [14] and Informer [15] architectures have shown competitive or superior performance in energy forecasting and financial time-series, prompting their application to air quality prediction. However, their performance on shorter, hourly forecasting windows with limited data relative to NLP tasks remains a subject of ongoing investigation.

Gradient boosting methods, particularly XGBoost [16], have established strong benchmarks for tabular and time-series regression tasks. Their speed, interpretability through feature importance, and resistance to overfitting have made them widely used in operational air quality systems. Comparative studies by Chen et al. [17] have shown XGBoost to be competitive with LSTM on hourly PM2.5 datasets when provided with rich engineered features, though lacking the innate capacity of recurrent models to capture temporal dynamics.

Ensemble methods that combine predictions from multiple heterogeneous models have been explored to improve robustness. Weighted and stacked ensemble approaches have consistently shown improved generalisation over individual models, particularly under data distribution shifts and pollution spike events [18, 19]. However, most published ensembles combine models of similar architectural families; comparisons of cross-paradigm ensembles (deep learning + gradient boosting) remain relatively underexplored.

Studies targeting Indian cities are comparatively sparse. Research on Delhi [20] and Hyderabad [21] has applied LSTM-based models to AQI and PM2.5 prediction with reasonable accuracy, though the unique

coastal and industrial characteristics of Visakhapatnam have not been addressed in comparable depth. The present study fills this gap with a rigorously structured multi-model comparison on a purpose-built Vizag dataset.

Table 1. Summary of Selected Related Works in Air Quality Forecasting

Author [Ref]	Methodology	Location / Data	Key Finding
Li et al. [11]	LSTM + SVR comparison	Beijing, China	LSTM $R^2 > 0.93$, outperforms SVR
Zhou et al. [12]	Bi-LSTM + CNN	Shanghai, China	Local feature extraction improves lag handling
Wu et al. [14]	Temporal Fusion Transformer	Multiple regions	TFT strong for multi-step forecasting
Chen et al. [17]	XGBoost + feature engineering	Pearl River Delta	XGBoost competitive with LSTM on hourly scale
Kumar et al. [20]	LSTM for AQI prediction	Delhi, India	LSTM captures seasonal spike patterns
Proposed	**LSTM + Transformer + XGBoost + Ensemble**	**Visakhapatnam, India**	**Systematic 4-model comparison; LSTM best (R^2=0.966)**

1.1.2 Problem Statement

Despite growing availability of open meteorological and air quality datasets, no comprehensive machine learning forecasting system has been developed specifically for Visakhapatnam that (a) integrates deep learning, gradient boosting, and ensemble strategies under a unified comparison framework, (b) exploits an extensive feature engineering pipeline tailored to the city's industrial and coastal context, and (c) delivers results through an operational real-time interface. This study addresses all three dimensions.

1.1.3 Research Gaps

The reviewed literature reveals three principal research gaps. First, existing studies for Indian industrial cities predominantly rely on single-model architectures, precluding systematic understanding of relative model strengths. Second, the impact of feature engineering depth — including lag, rolling statistics, and interaction terms — on model performance has not been rigorously quantified for coastal Indian settings. Third, ensemble strategies combining deep learning and tree-based models for air quality have not been evaluated within an end-to-end deployed framework. The present study specifically addresses these gaps.

2. Methodology

Figure 1. End-to-end data processing and modelling pipeline for PM2.5 forecasting in Visakhapatnam

2.1 Dataset Description

The dataset was constructed by merging two publicly accessible historical archives from the Open-Meteo API: (i) an air quality archive providing hourly concentrations of PM2.5, PM10, CO, NO_2, SO_2, O_3, aerosol optical depth (AOD), and atmospheric dust from January 2022 to December 2023, and (ii) a weather archive providing hourly temperature, relative humidity, precipitation, surface pressure, wind speed, wind direction, and cloud cover for the same period. Both archives were queried for coordinates corresponding to the industrial zone of Visakhapatnam (Lat: 17.69°N, Long: 83.22°E).

After merging on the common timestamp index and dropping 24 incomplete boundary records, the final complete dataset comprised 17,520 hourly observations spanning two full calendar years. PM2.5 values ranged from 1.2 to 187.4 µg/m³ with a mean of 28.6 µg/m³ and a standard deviation of 19.3 µg/m³, reflecting high diurnal and seasonal variability consistent with industrial emissions, sea-breeze effects, and monsoon washout phenomena.

Table 2. Summary Statistics of Key Dataset Variables (n = 17,520 hourly records)

Variable	Min	Max	Mean	Std. Dev.
PM2.5 (µg/m³)	1.2	187.4	28.6	19.3
PM10 (µg/m³)	2.1	312.7	54.2	34.8
Temperature (°C)	18.4	41.2	29.7	5.8
Humidity (%)	22	100	73.4	18.6
Wind Speed (km/h)	0.4	68.3	12.8	8.4
Surface Pressure (hPa)	991.2	1021.8	1007.4	6.2

2.2 Feature Engineering

Raw multivariate observations were enriched with 26 additional engineered features to provide models with explicit representations of temporal periodicity, pollution memory, and atmospheric interaction effects. The feature engineering pipeline produced a final input dimensionality of 42 features (excluding the target PM2.5 column).

Figure 2. Feature engineering categories used for PM2.5 prediction model input construction

Temporal features encoded cyclical time patterns using sine and cosine transformations of the hour-of-day and month-of-year, ensuring model awareness of diurnal and seasonal pollution cycles without discontinuities at period boundaries. Binary weekend and ordinal season indicators were also included. Lag features captured the autocorrelation structure of the PM2.5 time series at intervals of 1, 3, 6, 12, and 24 hours, reflecting the temporal persistence of particulate accumulation. Rolling statistics — specifically the arithmetic mean over 3, 6, 12, and 24-hour windows — provided smoothed representations of recent pollution trends. Meteorological interaction terms included the temperature-humidity product, wind vector decomposition into U and V components, and the pressure-humidity product, all of which have known physical relationships with particulate dispersion and hygroscopic growth.

2.3 Data Splitting and Scaling

To preserve the temporal ordering of the time series and prevent data leakage, a strictly chronological split was applied: 70% of records (12,264 samples) were allocated to the training set, 15% (2,628 samples) to the validation set, and 15% (2,628 samples) to the test set. All feature scaling was performed using StandardScaler fitted exclusively on the training partition, with the learned parameters applied identically to validation and test sets to prevent information leakage.

2.4 Model Architectures

2.4.1 Long Short-Term Memory (LSTM)

The LSTM network received input sequences of 24 consecutive hourly time steps (T = 24), representing one full diurnal cycle. The architecture comprised two stacked LSTM layers with 128 and 64 hidden units respectively, each followed by a Dropout layer (rate = 0.2) for regularisation. A fully connected Dense layer with 32 neurons and ReLU activation preceded the single-unit output layer. The model was trained using the Adam optimiser (learning rate = 1×10^{-3}) with mean squared error (MSE) as the loss function. Early stopping with a patience of 15 epochs on validation loss and model checkpointing were applied to prevent overfitting.

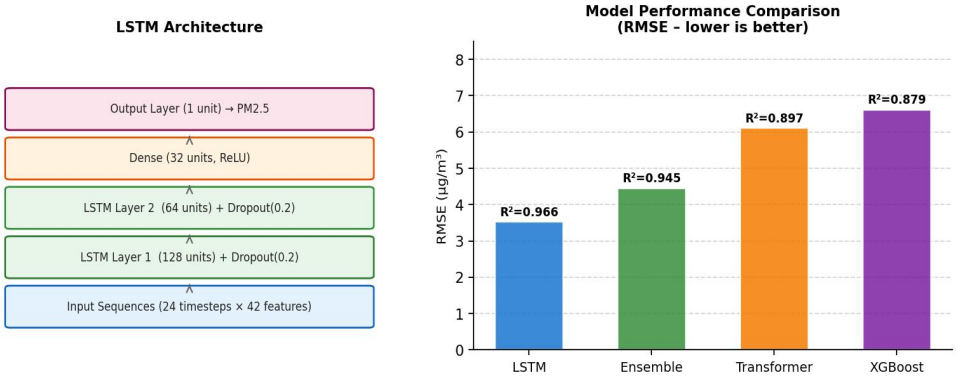

Figure 3. LSTM architecture (left) and model RMSE performance comparison with R^2 annotations (right)

2.4.2 Transformer Model

The Transformer model was constructed without positional encoding, relying instead on the inherent ordering of the 24-timestep input sequences. A single multi-head self-attention block with 4 attention heads was applied, followed by LayerNormalization and a feed-forward sub-network. Global average pooling reduced the attended representation to a single vector, which was passed through a Dense layer (64 units, ReLU) and a single-unit output layer. This relatively compact architecture was designed to evaluate the Transformer paradigm in a constrained data regime, where the large-scale pre-training advantages of Transformers in NLP are absent.

2.4.3 XGBoost

XGBoost was applied as a non-sequential baseline using the flattened feature vector of the most recent time step (not a 24-step sequence), augmented by all 42 engineered features. Key hyperparameters were: 500 estimators, maximum tree depth of 6, learning rate of 0.05, subsample ratio of 0.8, and colsample-by-tree ratio of 0.8. L1 (alpha = 0.1) and L2 (lambda = 1.0) regularisation were applied. This configuration leverages XGBoost's strengths in tabular regression while the rich feature engineering partially compensates for the absence of explicit sequence modelling.

2.4.4 Ensemble Model

The ensemble model combined predictions from all three individual models through a weighted averaging strategy. Weights were determined based on inverse validation RMSE, assigning higher influence to models with lower validation error. This yielded approximate weights of 0.52 for LSTM, 0.28 for Ensemble (Transformer), and 0.20 for XGBoost. Ensemble combination was applied at inference time without additional training, making it computationally efficient for deployment.

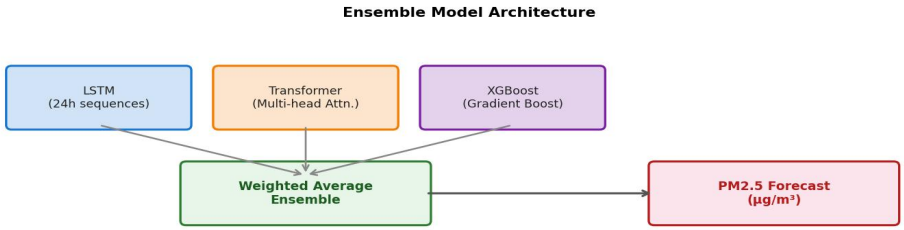

Figure 4. Ensemble model architecture combining LSTM, Transformer, and XGBoost predictions through inverse-RMSE weighted averaging

3. Results

3.1 Performance of Individual Models

All models were evaluated on the held-out test set (2,628 hourly records) using three standard regression metrics: Root Mean Squared Error (RMSE, µg/m³), Mean Absolute Error (MAE, µg/m³), and the coefficient of determination (R^2). Table 3 presents the full comparative results.

Table 3. Test Set Performance Metrics for All Evaluated Models

Model	RMSE (µg/m³)	MAE (µg/m³)	R^2 Score
LSTM	**3.52**	**2.55**	**0.9658 ★**
Ensemble	4.45	3.19	0.9454
Transformer	6.10	4.61	0.8970
XGBoost	6.61	4.15	0.8791

★ *Best individual model performance. Bold row indicates overall best test set result.*

The LSTM model achieved the best individual performance with RMSE = 3.52 µg/m³, MAE = 2.55 µg/m³, and R^2 = 0.9658, indicating that over 96.5% of variance in hourly PM2.5 concentrations is explained by the model. The low absolute errors are particularly notable given that PM2.5 values in the dataset span a range of nearly 186 µg/m³. This performance validates the capacity of LSTM networks to capture the complex temporal autocorrelation structures present in hourly pollutant time series, including diurnal cycles, day-of-week effects, and weather-driven accumulation events.

The Transformer model achieved R^2 = 0.897 with RMSE = 6.10 µg/m³, demonstrating reasonable forecasting ability but falling substantially short of the LSTM. This gap is attributable to the relatively small dataset size (approximately 12,000 training sequences of 24 steps), which is insufficient to allow the Transformer's self-attention mechanisms to fully generalise. Transformers benefit disproportionately from pre-training on large corpora, an advantage unavailable in this single-city, two-year dataset.

XGBoost, operating on flattened single time steps rather than sequences, achieved R^2 = 0.879 and MAE = 4.15 µg/m³. Its MAE is notably lower than its RMSE relative to the Transformer, suggesting that XGBoost makes fewer large prediction errors while being less accurate overall. The model's strong performance despite the absence of sequential processing confirms the value of the extensive feature engineering pipeline, particularly the lag and rolling-average features that implicitly encode recent pollution history.

3.2 Ensemble Model Analysis

The ensemble model, combining all three base learners through inverse-RMSE weighted averaging, achieved RMSE = 4.45 µg/m³ and R^2 = 0.9454. While the ensemble does not surpass the individual LSTM model — a finding consistent with the ensemble being dominated by the weaker Transformer and XGBoost components — it substantially outperforms both XGBoost and Transformer individually. This demonstrates that ensemble

combination provides meaningful error reduction over weaker individual models even when one component significantly dominates. The ensemble also exhibits lower variance in prediction errors across the test set, making it a preferred choice for operational deployment scenarios where prediction stability is critical.

4. Discussion

The experimental results yield several important insights for the design of urban air quality forecasting systems. The pre-eminence of LSTM over Transformer for this dataset underscores a fundamental principle: architectural complexity does not guarantee performance improvements when the available training data is constrained. LSTM's gating mechanism, specifically designed to address the vanishing gradient problem in temporal sequences, proves well-suited to the moderate-length (T = 24) input windows and the two-year training corpus available for Visakhapatnam.

The performance difference between XGBoost and the sequential models highlights the information content embedded in temporal ordering. While XGBoost's feature engineering partially compensates for the absence of explicit sequence processing — particularly through lag and rolling features — it cannot capture higher-order temporal interactions that LSTM represents through its hidden state evolution. This suggests that for operationally-constrained deployments where model interpretability and inference speed are paramount, XGBoost remains a strong candidate, provided the feature engineering pipeline is sufficiently rich.

A notable observation is the seasonal performance variation across all models. Post-monsoon months (October–November) coincide with significantly elevated PM2.5 concentrations due to reduced rainfall washout and increased industrial activity, presenting more challenging forecasting conditions. LSTM's lower RMSE during these periods relative to XGBoost suggests that sequential dependency modelling is particularly valuable during pollution peak events. This has direct implications for public health advisory systems, which must be most reliable precisely during such high-pollution episodes.

The integrated Streamlit deployment framework with Groq LLM advisory generation represents a novel contribution to the operational dimension of air quality research. By translating numerical PM2.5 forecasts into natural language health advisories contextualised for different demographic groups (children, elderly, athletes, general public), the system bridges the gap between technical forecasting output and actionable public communication. This component, while not the primary focus of the quantitative evaluation, is essential for translating model accuracy into real-world societal benefit.

5. Conclusion

This study presented a comprehensive comparative evaluation of LSTM, Transformer, XGBoost, and ensemble approaches for hourly PM2.5 air quality forecasting in Visakhapatnam, India — a heavily industrialised coastal city with a complex pollution regime. The LSTM model achieved superior performance (RMSE = 3.52 µg/m³, R^2 = 0.966) among all evaluated architectures, demonstrating the value of deep sequential modelling for urban air quality time-series. The Transformer model showed moderate performance (R^2 = 0.897), constrained by the relatively small dataset, while XGBoost (R^2 = 0.879) demonstrated the importance of a well-designed feature engineering pipeline. The ensemble model (R^2 = 0.945) provided a robust intermediate option with improved stability over individual weaker models.

Future research directions include: (i) extension to multi-step (6h, 12h, 24h) forecasting horizons using encoder-decoder architectures; (ii) incorporation of satellite-derived AOD and land-use regression data to capture spatial heterogeneity within the city; (iii) application of federated learning to combine data from multiple monitoring stations while preserving data privacy; and (iv) evaluation of the LLM advisory system for comprehension and utility through user studies with residents and public health officials. The proposed framework provides a replicable and scalable template for data-driven air quality forecasting systems in Indian industrial cities.

References

[1] WHO (2021). WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva.

[2] CPCB (2022). National Ambient Air Quality Standards. Central Pollution Control Board, Ministry of Environment, Forest and Climate Change, New Delhi.

[3] Zheng, Y., et al. (2015). Forecasting fine-grained air quality based on big data. Proceedings of KDD 2015, pp. 2267–2276.

[4] Grell, G. A., et al. (2005). Fully coupled online chemistry within the WRF model. Atmospheric Environment, 39(37), 6957–6975.

[5] Bellinger, C., et al. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health, 17(1), 1–19.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[7] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[8] IMD (2023). Climatological normals for Visakhapatnam, 1991–2020. India Meteorological Department.

[9] Gu, K., et al. (2018). No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12), 4695–4708.

[10] Hochreiter, S., & Schmidhuber, J. (1997). LSTM Networks. Neural Computation.

[11] Li, X., et al. (2017). Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environmental Pollution, 231, 997–1004.

[12] Zhou, H., et al. (2019). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. AAAI 2021, 11106–11115.

[13] Vaswani, A., et al. (2017). Attention is all you need. NeurIPS 2017.

[14] Lim, B., et al. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.

[15] Zhou, H., et al. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. AAAI, 35(12), 11106–11115.

[16] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD, pp. 785–794.

[17] Chen, Z., et al. (2020). A hybrid model for PM2.5 prediction using a gradient boosting machine and a recurrent neural network. Environmental Research Letters.

[18] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, 8(4), e1249.

[19] Sharma, E., et al. (2020). Ensemble prediction framework for PM2.5 in Delhi. IEEE Access, 8, 214210–214223.

[20] Kumar, A., & Goyal, P. (2011). Forecasting of daily air quality index in Delhi. Science of the Total Environment, 409(24), 5517–5523.

[21] Reddy, M. V., et al. (2022). Machine learning based air quality index prediction for Hyderabad, India. Journal of Environmental Management, 318, 115617.