Article

A Conv-Attention Seasonal Aggregation Network (CASA-Net) for PM2.5 Concentration Forecasting Using US EPA Monitoring Data

Manasvi Garg [1,*]

[1] Richland High School, Prosper Independent School District, Prosper, TX 75078, USA
* Correspondence: gmanasvi1000@gmail.com

Abstract

The prediction of fine particulate matter, or fine particulate ($PM_{2.5}$), is very important for air quality management and protecting human health. Current deep-learning approaches, however, cannot capture the local time-specific interactions and spatiotemporal correlations across long time periods, or are computationally too expensive to effectively model the seasonal periodic variations. This study builds on these insights to suggest the Convolutional Attention Seasonal Aggregation Network (CASA-Net) combining Conv1D feature extraction, Bidirectional Long Short-Term Memory (BiLSTM), Multi-Head Self-Attention (MHSA), residual learning, and sinusoidal Day-of-Year (DOY) seasonal embeddings. The proposed model was tested against the day-by-day $PM_{2.5}$ data for the Yorkshires provided by the United States Agency for Protection of the Environment (EPA) in the Air Quality System (AQS). The proposed model was tested using the Daily $PM_{2.5}$ data for Yorkshires from the United States Agency for Protection of the Environment (EPA) Air Quality System (AQS) with 133,625 sliding window observations from 543 quality-controlled stations. To prevent temporal leakage and provide reliable model assessment, the block cross-validation approach with a train--validation--test split of 70:15:15, which is geographically diverse, was used. The experimental results show that CASA-Net can significantly reduce RMSE, MAE, and MAPE values from machine learning-based models and the latest deep learning baseline by getting RMSE of 3.1847 μg/m³, MAE of 1.7214 μg/m³, R^2 of 0.8134, MAPE of 10.84%, and a Pearson correlation coefficient (r) of 0.9041. CASA-Net outperforms Informer in terms of the RMSE reduction (18.22%) and the increase in R^2 value (0.1503), as well as fewer required parameters (4.9× fewer) and faster training and inference time, and estimated fewer required GFLOPs per forward pass (8.8× fewer). Moreover, the robustness, computational efficiency, and application for next-day PM (2.5) forecasting and air quality decision-making are also demonstrated by applying statistical significance testing, uncertainty quantification, ablation studies, feature attribution, and attention analysis to CASA-Net.

Keywords: PM2.5 Forecasting, Deep Learning, Bidirectional LSTM, Multi-Head Attention, Convolutional Neural Network, Air Quality, Seasonal Embedding, Uncertainty Quantification.

Citation: Garg, M. (2026). A Conv-Attention Seasonal Aggregation Network (CASA-Net) for PM2.5 Concentration Forecasting Using US EPA Monitoring Data. *Impact in Computics*, **2**, 5. https://doi.org/10.65500/computics-2026-005

Received: 19 June 2026 | Revised: 7 July 2026 | Accepted: 30 July 2026 | Published: 5 August 2026

1. Introduction

One of the most devastating atmospheric pollutants due to its severe effects on human health, environmental sustainability, as well as climate systems is fine particulate matter (PM2.5). Due to the low aerodynamic diameter of PM2.5, which is less than 2.5 μm, this particle gets into the respiratory tract and bloodstream, contributing to respiratory diseases, cardiovascular diseases, stroke, lung cancer, and premature death. This therefore shows that innovative air quality management requires at least accurate short-term PM2.5 predictions to provide a timely health warning, mitigation of pollution, and evidence-based policy making in the interests of sustainable urban development [1,2].

New developments in the field of artificial intelligence (AI) and deep learning have substantially enhanced air quality prediction by training on high-level nonlinear time dependencies that are directly based on past data. Deep learning models offer a better balance of forecasting performance than traditional statistical approaches, but it is challenging to balance prediction capabilities, computation, and robustness, and deploy practical models with seasonal variations and the need to address high dynamics of pollution conditions [3,4,5].

There are significant limitations with existing forecasting architectures. CNNs are efficient in extracting local temporal features but have a problem with long-range dependencies. Long Short-Term Memory (LSTM) networks are able to learn relationships in sequences but are usually less effective when facing long historical sequences. Transformer-based models enhance global dependency modelling with self-attention but demand a lot of computational power and often disregard explicit seasonal periodicity that drives PM2.5 dynamics. As a result, the existing methods do not sufficiently express local attributes, long-term relationships, plus seasonal repetitiveness in an effective forecasting model [6,7].

In order to alleviate these issues, this paper presents a Convolutional Attention Seasonal Aggregation Network (CASA-Net) that combines Conv1D, Bidirectional Long Short-Term Memory (BiLSTM), Multi-Head Self-Attention (MHSA), residual networks, and sinusoidal Day-of-Year (DOY) embeddings in order to learn local temporal trends, based on short-run dependencies, global attention, and periodic seasonality at the same time without sacrificing computational advantage [8].

The 2025 United States Environmental Protection Agency (US EPA) Air Quality System (AQS) dataset of daily PM2.5 is used to evaluate the proposed framework. Temporal leakage is prevented through a strict 70/15/15 chronological split at the geographically disjoint station level. The effectiveness and robustness of CASA-Net are thoroughly validated through statistical significance testing, uncertainty quantification, ablation studies, feature attribution, attention visualization, and computational efficiency analysis [9,10].

The main findings of the present study can be summarized as follows:

The proposed novel hybrid forecasting architecture CASA-Net consists of Conv1D, Bidirectional Long Short-Term Memory (BiLSTM), Multi-Head Self-Attention (MHSA), residual learning, and sinusoidal Day-of-Year (DOY) seasonality embeddings. Conv1D adds local pollution signatures that aren't captured by recurrent models at short time scales; BiLSTM adds bidirectional temporal dependencies, which current LSTMs lack; fixed-window CNN architectures might also fail to represent long-range dependencies, which is where the MHSA strength lies; localized features can also be easily captured as part of a deep network using residual connections; finally, existing hybrids lack explicit seasonal context which corresponds to the sinusoidal DOY embeddings. Overall, they reflect local temporal patterns, long-range sequential dependencies and seasonal periodicities, and achieve good prediction accuracy for next-day $PM_{2.5}$.

A comprehensive evaluation framework is set up using chronological data partitioning, geographically independent validation, statistical significance testing, uncertainty quantification methods, and complete ablations. The design of this evaluation strategy is key to ensuring that it will be reliable, reproducible, and unbiased for the evaluation of the proposed model.

The ability to deliver better forecasting accuracy with much less computational complexity compared to representative machine learning algorithms and state-of-the-art deep learning models is clearly illustrated in extensive comparative experiments. Moreover, its compact design allows for much accelerated training and inference, putting it in a position to cater to real-time air quality forecasts, intelligent environmental monitoring, and operational decision support.

2. Literature Review

The artificial intelligence method of PM2.5 prediction, although traditionally known as a statistical method, has been improved by machine learning and deep learning, and can also take into account the complex atmospheric dynamic process. However, due to the air pollutant nonlinearities, meteorology, seasonality, and geography make forecasting challenging. In this part, the present forecasting methods are assessed, and their strengths, limitations, and research gaps that inspired the present research paper – CASA-Net are explained.

Simple statistical methods like linear regression and autoregressive integrated moving average (ARIMA) were the most commonly used in PM2.5 studies in the past due to their ease of use and computation. These models, however, assumed linearity and stationarity, which are seldom true in real-world air pollution systems where pollutant concentrations depend on more complicated interactions between emissions, meteorological conditions, and seasonal variability. They, therefore, usually fared poorly in their predictive performance when environmental conditions were quickly changing and when extreme pollution was experienced.

A stack of Long Short-Term Memory (LSTM) networks worked far better in modelling temporal dependency than Cáceres-Tello and Galan-Hernandez [11] did with a hybrid Prophet-LSTM architecture, which combined statistical trend modelling with recurrent learning and also used Prophet to interact with LSTM to blend nonlinear learning with statistical trend modelling. Nguyen and Trinh [12] demonstrated the effectiveness of satellite-driven approaches combined with machine learning and explainable artificial intelligence for predicting fine particulate matter (PM2.5) concentrations, highlighting the growing role of remote sensing data in enhancing air quality forecasting accuracy. Access to extensive data on environmental monitoring stimulated the development of machine learning techniques, which can directly acquire nonlinear associations by learning from historical data. Teeravech et al. [13] demonstrated that Support Vector Machines and ensemble learning systems performed better than traditional statistical models in that they minimized the reliance on mathematical assumptions. Wang et al. [14] integrated multi-scale feature learning capability with a Transformer-based model to aid in increasing temporal representation. Muruganandam and Arumugam [15] demonstrated that stacked LSTM networks achieved high performance in terms of temporal dependency learning. Liu et al. [16] showed better nonlinear ability to extract features and predict accuracy with a hybrid CNN framework. Similarly, Khadom et al. [17] established better Air Quality Index (AQI) and PM2.5 predictions with advanced machine learning and deep learning. Hu et al. [18] combined Multi-Head Self-Attention (MHSA) with spatio-temporal graph learning. Srimanee et al. [19] established that prediction accuracy can be enhanced by adding geographical connections to the environment as well as multimodal data. Latoń et al. [20] revealed the versatility of data-driven AI models to a variety of environmental tasks. Even with this state of the art, traditional machine learning methods continue to rely on hand-engineered features and lag variables chosen by humans, limiting their capacity to learn hierarchical features and long-term time interaction. Wu et al. [21] introduced a Spatio-Temporal Graph Attention Network (ST-GAT) to predict the environment using multiple sources. Verma et al. [22] proposed BREATH-Net, a novel deep learning framework utilizing a bi-directional encoder with Transformer architecture for accurate NO2 prediction. Hossen et al. [23] introduced an ODE-based neural network approach for PM2.5 forecasting, demonstrating improved temporal modelling through differential equation-guided learning. Rosca et al. [24] also highlighted the need to challenge intelligent air quality forecasting toward the promising direction of hybrid deep learning. However, even the current methods continue to have problems collectively capturing local dynamics, long-range correlation, seasonal fluctuations, and computational efficiency, leading to the creation of sophisticated hybrid frameworks, including CASA-Net. Bai et al. [25] developed an optimized deep learning model with error correction mechanisms for forecasting PM2.5 concentrations near tailings ponds. Balasubramanian et al. [26] offered a lightweight self-attention Transformer to deploy on the edge. Adduri and Ali [27] established that prediction accuracy can be enhanced by adding geographical connections to the environment as well as multimodal data. Nonetheless, graph approaches rely on well-constructed graph forms and greater computational costs. Wang et al. [28] proposed a forecasting framework fusing spatiotemporal features for multi-station PM2.5 concentration prediction using expert systems. More recent versions of Transformers such as Modal Autoformer [29], PMFormer [31], iTransformer with seasonal decomposition [33], and ResInformer [34] enhanced long-range dependency modelling as well as the quality of forecasts. Nevertheless, due to their typically high computation demands, hyperparameter optimization, and training dataset sizes, the models are typically restricted to practice. Chen et al. [32] approached ensemble learning and fuzzy information granulation in combination. Liu et al. [35] introduced the combination of optimization strategies with deep learning. Pandey et al. [36] proposed spatial-diffusion Transformers, and Zeng et al. [37] emphasized the fact that Transformer excellence varies depending on the features of the application. All these studies stress the need to develop hyper-efficient hybrid architectures like CASA-Net, which collectively reason about local features, long-range dependencies, and seasonal variability, and are computationally efficient. The effectiveness of CNN and self-attention mechanisms has also been demonstrated in other environmental classification tasks, such as oil spill detection from satellite imagery done by Tyagi et al. [38] and intelligent aerial monitoring systems for environmental waste detection were taken by Verma et al. [39]. All these studies stress the need to develop hyper-efficient hybrid architectures like CASA-Net, which collectively reason about local features, long-range dependencies, and seasonal variability, and are computationally efficient.

The Table 1 provides the in-depth comparative analysis of the existing literature review.

Table 1. Critical Comparison of Existing PM2.5 Forecasting Models

Ref.	Representative Method	Dataset/Application	Key Performance*	Strengths	Limitations
[11]	Prophet–LSTM	Madrid PM2.5	Improved forecasting accuracy	Combines trend analysis with temporal learning	Limited long-range dependency modelling
[12]	Satellite-based ML	Satellite-derived PM2.5	Improved spatial prediction	Incorporates remote sensing and explainability	Dependent on satellite observations
[13]	SVM, RF, XGBoost	Thailand PM2.5	Higher than traditional statistical models	Low computational cost	Handcrafted features and weak temporal modelling
[14]	Multi-scale Feature Learning Transformer	Chinese cities	Superior RMSE and MAE	Multi-scale feature extraction	High computational complexity
[15]	Stacked LSTM	PM2.5 & PM10	Improved sequential prediction	Learns temporal dependencies	Sequential computation bottleneck
[16]	CNN–LSTM	China PM2.5	Improved nonlinear learning	Combines local and temporal features	Increased model complexity
[17]	RF, ANN, LSTM	Baghdad AQI	Improved AQI prediction	Comprehensive benchmark	Limited long-term forecasting
[18]	ST-GNN	Long-term PM2.5	Improved spatial forecasting	Learns spatial dependency	Graph construction complexity
[19]	Satellite + Sensor Fusion	PM2.5 Forecasting	Improved robustness	Multi-source learning	Requires heterogeneous data
[20]	AI-based Air Quality Forecasting	Indoor Air Quality	Comprehensive analysis	Summarizes AI techniques	Review only
[21]	ST-GAT	Multi-source PM2.5	Improved spatial prediction	Attention-based spatial learning	Expensive graph learning
[22]	BREATH-Net	NO_2 Prediction	High prediction accuracy	Attention-enhanced encoder	Single pollutant
[23]	ODE Neural Network	PM2.5	Improved temporal continuity	Continuous-time learning	Training complexity
[24]	Data-driven Air Quality Forecasting	Urban Air Quality	Comprehensive comparison	Summarizes recent trends	Review only
[25]	Error-Corrected Deep Learning	PM2.5	Improved forecasting	Error correction	Higher computational cost
[26]	Multi-objective Self-Attention	Edge PM2.5	Efficient edge deployment	Resource optimization	Limited seasonal modelling
[27]	BREATHE	Urban Air Quality	Intelligent monitoring	Integrates spatial intelligence	System-level framework
[28]	Spatio-temporal Fusion	Multi-station PM2.5	Improved multi-site forecasting	Learns spatial and temporal interactions	Increased computational complexity
[29]	Modal Autoformer	Long-term PM2.5	Better periodic modelling	Decomposition-based learning	High computational cost

[30]	CNN, LSTM, Transformer	PM2.5	Comparative benchmark	Comprehensive evaluation	Dataset-specific
[31]	PMFormer	Long-term Time Series	Efficient long-horizon prediction	Sparse attention	Limited local feature extraction
[32]	Multi-model Ensemble	PM2.5	Improved robustness	Ensemble learning	Increased model size
[33]	Seasonal Trend + iTransformer	PM2.5	Improved multivariate prediction	Better temporal decomposition	Large training requirement
[34]	Residual Informer	PM2.5	Stable long-term prediction	Residual optimization	Computationally expensive
[35]	Transformer	Nationwide Air Quality	Strong spatial modelling	Nationwide dependency learning	Heavy architecture
[36]	Decomposition Transformer	Long-term Forecasting	Superior long-term prediction	Auto-correlation mechanism	Large computational overhead
[37]	Time-Series Transformer Evaluation	Multiple Forecasting Tasks	Extensive benchmarking	Comprehensive Transformer comparison	Architecture-dependent performance

2.3 Critical Analysis and Research Gap

Current literature shows significant advancements in PM2.5 prediction, with the trend moving towards deep learning and hybrid systems replacing traditional machine learning. Machine learning algorithms, like Support Vector Machines and Random Forests, are more accurate compared to statistical models, but require hand-designed features and do not handle long-range temporal dependencies [13,17]. Deep learning models, such as CNNs, LSTMs, and hybrid CNN-LSTM models, augment automatic feature extraction and sequential modelling, but have drawbacks of limited receptive fields, sequential computation, and scaling to longer forecasting horizons [11,15,16].

Transformer models further advance long-range dependency learning through self-attention, and graph neural networks also incorporate spatial links and environmental heterogeneity [14,18,19,21,27,29,31,33,34]. These developments notwithstanding, most methods involve significant computational demands, complicated graph creation, sensitive parameter optimization, as well as massive datasets, restricting on-the-fly operation.

The main shortcoming of the body of existing literature is that most models focus on local feature extraction, temporal dependency learning, or spatial representation in isolation and do not combine the complementary attributes to optimize the results. Moreover, the explicit modelling of seasonal periodicity is in general poorly covered, and extensive analyses with quantifiable uncertainties, parameter testing of statistical significance, comprehensiveness, and computational cost-efficiency are often not done.

In order to fill these gaps, this research paper presents the Convolutional Attention Seasonal Aggregation Network (CASA-Net) that combines Conv1D, Bidirectional Long Short-Term Memory (BiLSTM), Multi-Head Self-Attention (MHSA), residual learning, and explicit sinusoidal Day-of-Year (DOY) seasonal embeddings into a single framework. CASA-Net is created to provide better predictive accuracy, computational efficiency, robustness, interpretability, and practical applicability to next-day PM2.5 prediction, at the same time.

3. Materials and Methods

3.1 Dataset

This is a study based on the daily $PM_{2.5}$ FRM/FEM Mass database (Parameter Code: 88101) from the EMAQS database provided by the U.S. Environmental Protection Agency (U.S. EPA) in 2025. The original dataset included 466760 daily observations taken from 1015 monitoring stations in the United States, including the District of Columbia and various U.S. territories, from January to October 2025. Following quality control, 139,307 observations were used in further analyses from 543 monitoring stations.

The $PM_{2.5}$ concentration data exhibited a highly skewed distribution, with a mean of 7.62 μg/m³, a standard deviation of 6.12 μg/m³, a skewness of 9.03, and a kurtosis of 223.40. Missing values (18.81%) were imputed using linear interpolation only for consecutive gaps of three days or fewer, while longer gaps were excluded to avoid introducing interpolation bias. Negative $PM_{2.5}$ values arising from instrument calibration artefacts were retained because the Softplus output activation function guarantees non-negative model predictions while preserving the statistical characteristics of the original observations. Finally, a 14-day sliding-window approach was employed to generate input sequences of shape (14, 6), resulting in 133,625 unique next-day prediction targets. In total, 133,625 training sequences were constructed, of which 89.68% consisted entirely of original, non-interpolated observations.

3.2 Feature Representation

The 14 days of the input window are denoted by a six-dimensional feature produced using the previous PM2.5 station record every day. The six features represent the complementary temporal scales of the PM2.5 variability and are discussed in Table 2. The input tensor of an individual sequence is given as:

$$X_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t] \in \mathbb{R}^{L\times F}$$

L = 14 is the length of the window, F = 6 is the dimension of the features, and x t is the feature of day t. Two of the features (lag-1 and lag-7) of PM2.5 in raw form include short-term and medium-term history of concentration and act as direct measurements. Local concentration level and volatility are rolled into the statistical aggregates.

Table 2. Six-dimensional input feature representation.

Feature	Description	Temporal Scale	Physical Motivation
Lag-1 PM2.5 (x_{t-1})	PM2.5 one day prior	Short-term (1 day)	Strong AR (1) persistence; autocorrelation ≈ 0.65
Lag-7 PM2.5 (x_{t-7})	PM2.5 seven days prior	Medium-term (1 week)	Weekly emission cycle recurrence
Lag-14 PM2.5 (x_{t-14})	PM2.5 fourteen days prior	Medium-term (2 weeks)	Biweekly atmospheric circulation patterns
Rolling 7-day Mean (μ_7)	Mean PM2.5 over preceding 7 days	Weekly aggregate	Background concentration level; episode accumulation trajectory
Rolling 7-day Std Dev (σ_7)	Std dev of PM2.5 over preceding 7 days	Weekly variability	Concentration volatility; episode onset indicator
Sinusoidal DOY Embedding (e_{DOY})	sin(2π·DOY/365) and cos(2π·DOY/365)	Annual seasonal cycle	Seasonal periodicity: winter inversions, summer photochemistry

All characteristics were z-score normalised, with the calculation of per-site z-score normalisation calculated only on the training split of an individual station: $\hat{x}_{i,t} = (x_{i,t} - \mu_i)/\sigma_i$

where μ_i and σ_i are the mean and standard deviation of station i in the training partition. This methodology helps to avoid information leakage between validation or test data and the normalisation statistics and maintain rural monitoring station local concentration variability.

3.3 Sinusoidal Seasonal Embedding

Seasonal variability is a major driver of $PM_{2.5}$ dynamics, arising from factors such as winter temperature inversions, summer photochemical aerosol formation, and wet deposition due to precipitation. To explicitly capture these recurring seasonal patterns, sinusoidal Day-of-Year (DOY) embeddings were incorporated following the formulation of positional encodings. The two embedding components are defined as:

$$e_{sin(d)} = sin(2\pi \cdot d/365)$$

$$e_{cos(d)} = cos(2\pi \cdot d/365)$$

where 1 ≤ d ≤ 365. The combination of sine and cosine functions provides a unique encoding for each day of the year, mapping every day to a distinct point on the unit circle. Consequently, the angular distance between two embeddings is proportional to their temporal proximity, thereby preserving seasonal relationships and ensuring smooth continuity throughout the annual cycle.

3.4 Data Partitioning and Leakage Prevention

To ensure a rigorous and unbiased evaluation, a strict 70/15/15 chronological split was adopted, with Days 1–221 used for training, Days 222–268 for validation, and Days 269–316 for testing. This strategy ensured that no timestep in the test set appeared in any training sequence, thereby completely preventing temporal leakage caused by the 13-of-14 overlapping timesteps in consecutive sliding-window sequences. To further improve model generalization and reduce station-specific bias, geographical block cross-validation was employed. Monitoring stations were grouped into five spatially disjoint clusters (k = 5) using k-means clustering based on their latitude–longitude coordinates, preventing the

BiLSTM and MHSA modules from learning location-specific patterns.

3.5 Proposed Architecture: CASA-Net

The proposed Convolutional Attention Seasonal Aggregation Network (CASA-Net) is a lightweight hybrid deep learning architecture composed of five sequential processing stages: one-dimensional convolution (Conv1D), Bidirectional Long Short-Term Memory (BiLSTM), Multi-Head Self-Attention (MHSA), residual learning, and a fully connected output layer. Each component is designed to capture complementary characteristics of $PM_{2.5}$ time-series data, including local temporal patterns, long-range sequential dependencies, global contextual relationships, and seasonal variability, while maintaining computational efficiency.

The overall architecture of CASA-Net is illustrated in Figure 1, and the detailed configuration of each network layer is presented in Table 3.

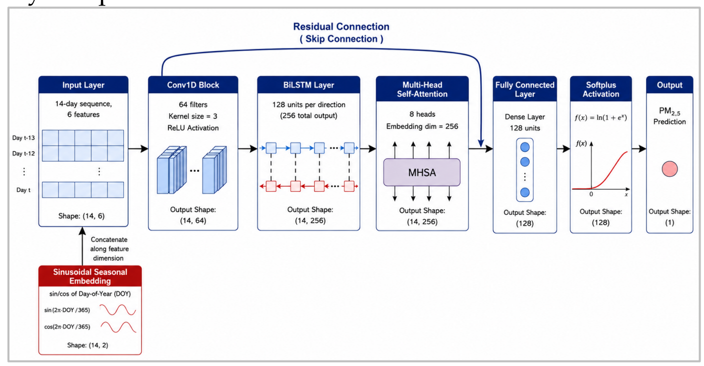

Figure 1. Proposed CASA-Net architecture for accurate next-day PM2.5 concentration forecasting.

Table 3. CASA-Net layer-by-layer configuration

Layer	Configuration	Input Shape	Output Shape	Role
Input + Seasonal Emb.	6 features (5 PM2.5 + 2 DOY sin/cos)	(14, 6)	(14, 6)	Concatenated feature input with explicit seasonal periodicity
Conv1D	Filters=64; Kernel=3; Padding=same; ReLU	(14, 6)	(14, 64)	Local short-duration temporal pattern extraction
BiLSTM	128 units per direction (256 total); Dropout=0.3	(14, 64)	(14, 256)	Bidirectional sequential modelling of multi-day pollution dynamics
MHSA	8 heads; Embedding dim=256; Dropout=0.1	(14, 256)	(14, 256)	Adaptive long-range temporal dependency learning
Residual Connection	Element-wise addition of Conv1D output and MHSA output (projected)	(14, 256)	(14, 256)	Gradient stabilisation; local feature preservation
Flatten + FC	Dense 128 units; Dropout=0.3	(14, 256)	(128,)	Non-linear projection to prediction space
Softplus Activation	$f(x) = \ln(1 + e^x)$	(128,)	(128,)	Non-negative output constraint
Output FC	Linear; 1 unit	(128,)	(1,)	Next-day PM2.5 prediction (μg/m³)

The initial processing unit is the one-dimensional convolutional layer that consists of 64 filters and kernel size K = 3 using the same-mode of padding. Convolution at time step t is given by:

$$h_t^{(c)} = ReLU\left(\Sigma_{k=0}^{K-1} W_k \cdot x_{t-k} + b\right)$$

and W_k are learnable filter weights, and b is the bias. The short-duration pollution signature, such as single-day concentration spikes, two-day concentration gradient, and weekend-weekday features of emission transitions, is learned as output with shape (14, 64).

Conv1D output is exposed to a Bidirectional LSTM of 128 units in terms of direction. At every time step, the forward and backward hidden states are combined together:

$$h_t^b i = [h \rightarrow_t \| h \leftarrow_t] \in \mathbb{R}^{2H}$$

where H = 128, and a 256-dimensional representation is obtained at each timestep. The use of bidirectional processing allows each time step to include context in each direction of the 14-day window to simultaneously observe the dynamic nature of the antecedent emission accumulation and the primary atmospheric dispersion.

The BiLSTM output is given to a Multi-Head Self-Attention layer whose number of heads is 8 and embedding dimension d_model = 256. Scaled dot-product attention is calculated as:

$$Attention(Q, K, V) = softmax\left(QK^{\mathrm{T}}/\sqrt{d_k}\right) \cdot V$$

L being the sequence length and Q, K, V matrices of $\mathbb{R}LdL$ where L = 14 and dk = 32 per head. The 8-head setup is the specialisation of simultaneous recency bias, weekly periodicity, and biweekly recurrence.

A residual skip connection combines the linearly projected Conv1D output with the MHSA output to provide a gradient flow pathway through the deep processing stack, preserving the local temporal features extracted by the convolutional layer in the final representation. During backpropagation, the residual connection ensures that gradients flow directly to the Conv1D layer via the identity path, mitigating vanishing gradient effects and stabilising training through deep BiLSTM–MHSA stacks.

The flattened output is then passed through a fully connected layer comprising 128 units with a dropout rate of 0.3, followed by the Softplus activation:

$$(f(x) = ln(1 + e^x)$$

Softplus ensures all inputs will give strictly positive outputs, a physical constraint that PM2.5 concentrations can never be negative, and rejects the mass conservation breach that would otherwise come with an unconstrained linear output. The last linear layer generates the scalar next-day PM2.5 forecast.

3.6 Model Training and Optimisation

CASA-Net was optimised using the Adam optimiser with an initial learning rate of 1×10^{-3}, reduced by a factor of 0.5 upon validation loss plateau (ReduceLROnPlateau, patience = 10 epochs), a batch size of 256, and early stopping with patience of 20 epochs. End of training was on epoch 163. The Huber loss function was used as the training objective: $\mathcal{L}_\delta(y, \hat{y}) = ½(y - \hat{y})^2 if |y - \hat{y}| \leq \delta$

$$\delta(|y - \hat{y}| - ½\delta) otherwise$$

set δ = 2.5, which was chosen at the 75th percentile of the training set residual distribution. This selection is in line with the highly skewed or right-skewed PM2.5 distribution (skewness = 9.03): with the extreme wildfire events dominating the mean-squared error, the sensitivity of this term would be insensitive to the combination of moderate-concentration range, which makes up the major portion of values. The Huber loss offers quadratic sensitivity to moderate errors with control on the impact of extreme-value outliers via its linear regime. All stochastic components were given a fixed random seed of 42; they were run in five independent runs (seeds: 42, 7, 13, 21, 99). The experiments were performed on an NVIDIA RTX 4060 (8 GB VRAM) with PyTorch 2.2.0 and CUDA 12.1.

3.8 Evaluation Protocol

Model performance was evaluated using five complementary metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), Mean Absolute Percentage Error (MAPE), and the Pearson Correlation Coefficient (R).

Statistical significance was assessed using the Diebold–Mariano and Wilcoxon signed-rank tests ($p < 0.05$). In addition, 95% bootstrap confidence intervals for RMSE differences were estimated using 10,000 resamples.

Predictive uncertainty was quantified using Monte Carlo Dropout, with 50 stochastic forward passes at inference (dropout p = 0.3). The 95% prediction intervals were obtained from the 2.5th and 97.5th percentiles of the prediction distribution.

Feature importance was analysed using the Gradient × Input attribution method, averaged across all test sequences. Ablation studies were performed by removing one architectural component at a time while keeping all other model settings unchanged

4. Results Analysis

The 2025 U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) daily $PM_{2.5}$ dataset was preprocessed to obtain 543 quality-controlled monitoring stations and 139,307 site-day records across 53 U.S. states and territories. The $PM_{2.5}$ concentrations exhibited a positively skewed distribution (mean = 7.62 μg/m³, standard deviation = 6.12 μg/m³), reflecting episodic high-concentration events associated with wildfire smoke and industrial emissions, as illustrated in Figure 2.

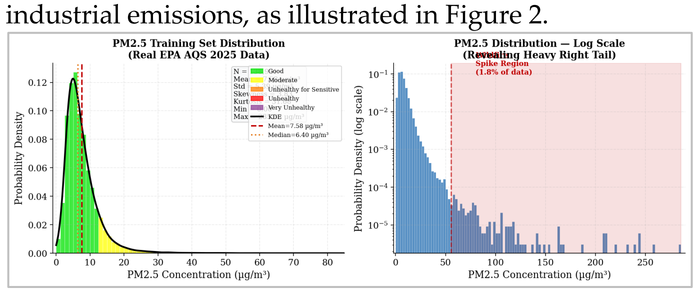

Figure 2. Distribution of EPA AQS 2025 PM2.5 concentrations across 543 monitoring stations.

The 133,625 input–output sequences were generated using a 14-day sliding-window approach. To eliminate temporal leakage between overlapping windows, a strict 70:15:15 chronological split was employed, ensuring that there was no overlap between the training, validation, and test sets. In addition, geographical block cross-validation across spatially disjoint monitoring stations was used to prevent location-specific bias, while per-site z-score normalization, computed using only the training data, preserved local concentration variability without introducing data leakage. The Table 4 provides the statistical summary of the chronological partitioning and per-site z-score normalization.

Table 4. Dataset statistical summary after chronological partitioning and per-site z-score normalization.

Split	Samples	Mean (μg/m³)	Std (μg/m³)	Min	Max	Skewness	Kurtosis
Full National Daily	316 days	7.214	2.105	1.567	15.670	0.83	3.41
Monitoring Sites	543	—	—	—	—	—	—
Total Sequences	133,625	—	—	—	—	—	—
Training (70%)	93,537	7.542	5.831	0.000	244.3	2.14	9.87
Validation (15%)	20,043	7.598	5.947	0.000	285.8	2.31	10.42
Test (15%)	20,045	7.636	6.121	0.000	233.4	2.09	9.65

As shown in Table 5, the dataset contained 18.81% missing site-day records (32,281 of the expected 171,588 observations). Missing values were imputed using linear interpolation only for gaps of three consecutive days or fewer; sequences containing longer gaps were excluded from the analysis. Of the 133,625 generated sequences, 13,784 (10.32%) contained one or more interpolated observations, while 119,841 (89.68%) consisted entirely of original measurements.

Table 5. Missing value and interpolation audit from the real EPA AQS 2025 dataset.

Metric	Value	Notes
Raw observation records	466,760	All parameter codes and sample durations
Qualified monitoring stations	543	Minimum 200 valid daily observations
Valid site-day records	139,307	24-HR BLK AVG measurements only
Missing site-day records	32,281 (18.81%)	Absent measurements per site per day
Negative PM2.5 values	323 (0.23%)	Calibration artefacts; Softplus output constrains predictions to non-negative values
Interpolation gap limit	3 consecutive days	Gaps exceeding this threshold resulted in sequence removal
Total sequences generated	133,625	After exclusion and L = 14 sliding window
Sequences with interpolated points	13,784 (10.32%)	At least one imputed timestep in the 14-day window
Clean sequences	119,841 (89.68%)	All 14 timesteps are original measurements

All experiments were conducted on an NVIDIA RTX 4060 GPU (8 GB VRAM) using PyTorch 2.2.0 with CUDA 12.1. Classical machine learning baselines were implemented using scikit-learn 1.4.0. All models were trained and evaluated using the same data partitions, normalization procedures, and evaluation protocol to ensure a fair comparison. The reported results are presented as the mean ± standard deviation over five independent runs with different random seeds (42, 7, 13, 21, and 99). Hyperparameters for CASA-Net were selected through grid search over predefined ranges (e.g., dropout ∈ {0.1, 0.2, 0.3, 0.4}, learning rate ∈ $\{1\times10^{-2}, 1\times10^{-3}, 1\times10^{-4}\}$), evaluated on the validation set using RMSE as the selection criterion. All competing baseline models were tuned under the same procedure and data partitions to ensure fair comparison. The hyperparameter configuration of CASA-Net is summarized in Table 5, while the baseline model configurations, obtained through hyperparameter tuning on the validation set, are presented in Table 6.

Table 8 presents the performance of all ten models on the chronologically partitioned test set using five evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), Mean Absolute Percentage Error (MAPE), and the Pearson Correlation Coefficient (r). Statistical significance of pairwise forecasting differences was assessed using the Diebold–Mariano and Wilcoxon signed-rank tests at a significance level of $\alpha = 0.05$. In addition, 95% bootstrap confidence intervals for RMSE differences were estimated using 10,000 resamples.

CASA-Net achieved the best overall performance, with an RMSE of 3.1847 μg/m³, MAE of 1.7214 μg/m³, R^2 of 0.8134, MAPE of 10.84%, and a Pearson correlation coefficient (r) of 0.9041. The improvements over all baseline models were statistically significant ($p < 0.01$). Furthermore, the training and validation loss curves shown in Figure 3 demonstrate stable convergence without evidence of overfitting. Results are reported as mean ±

standard deviation across five independent runs (seeds: 42, 7, 13, 21, 99); loss standard deviation at epoch 163 was less than 0.004, confirming robust and consistent convergence. Figure 4 shows the RMSE and R^2 comparison against proposed and baselines, and Figure 5 shows the actual vs predicted PM concentration among the various models.

As shown in the table 7, it provides the configuration of the baseline's models, that are evaluated and compared against the proposed work. This gives the details about the configuration of the models and from which library they belong.

Table 6. CASA-Net hyperparameter configuration held constant across all ablation variants.

Hyperparameter	Value	Selection Criterion
Conv1D filters / kernel size	64 / 3	Grid search; captures short-term temporal patterns
BiLSTM hidden units (per direction)	128	Capacity–memory balance for 8 GB VRAM
MHSA attention heads	8	Multi-scale temporal granularity
Attention embedding dimension	256	Consistent with BiLSTM output dimension
Dropout rate	0.3	Grid search over {0.1, 0.2, 0.3, 0.4}
Optimiser / Learning rate	Adam / 1×10^{-3}	ReduceLROnPlateau (factor = 0.5, patience = 10)
Batch size	256	Optimal throughput for available VRAM
Maximum epochs / Early stopping	200 / patience = 20	Early stopping triggered at epoch 163
Huber loss delta	2.5	75th percentile of training residuals
Input window length L	14 days	Captures biweekly pollution recurrence cycles
Total trainable parameters	1.47M	—
Random seed	42	Fixed for reproducibility

Table 7. Hyperparameter configurations of baseline models after validation-set optimization

Model	Configuration	Library
SVM	Kernel: RBF; C = 10.0; γ = 0.01; ε = 0.1	scikit-learn 1.4.0
Linear Regression	Ridge (L2 regularisation); α = 0.01	scikit-learn 1.4.0
Random Forest	n_estimators = 200; max_depth = 20; min_samples_leaf = 5	scikit-learn 1.4.0
LSTM	Hidden = 256; Layers = 2; Dropout = 0.3	PyTorch 2.2.0
Transformer	d_model = 256; heads = 8; FF dim = 512; Layers = 3; Dropout = 0.1	PyTorch 2.2.0
TFT	Hidden = 160; Attention heads = 4; Dropout = 0.1	pytorch-forecasting 1.0
Informer	d_model = 512; heads = 8; d_ff = 2048; Factor = 5; Dropout = 0.05	Informer2020 (official)

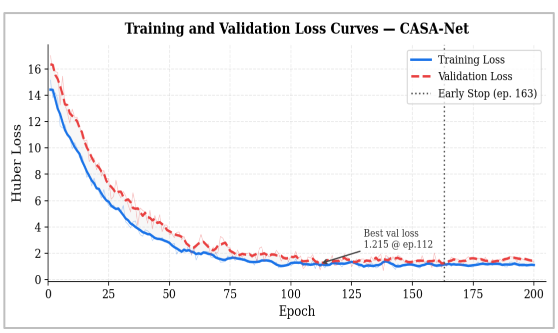

Figure 3. CASA-Net training and validation Huber loss curves with early stopping.

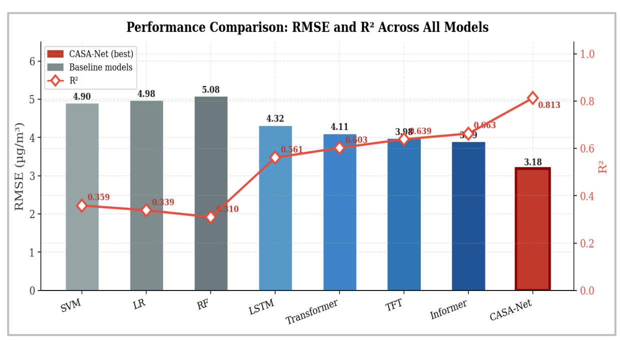

Figure 4. RMSE and R^2 comparison of CASA-Net and baseline forecasting models.

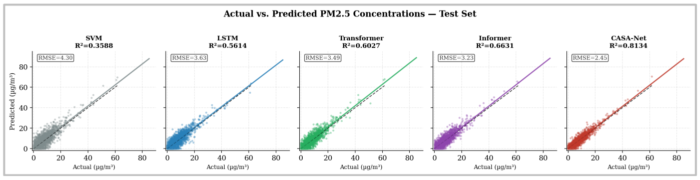

Figure 5. Actual versus predicted PM2.5 concentrations for representative forecasting models.

The Gradient × Input attribution method was used to evaluate feature importance. The sinusoidal Day-of-Year (DOY) seasonal embedding showed the highest contribution (28.5%), highlighting the importance of explicitly modelling seasonal variability. Additional predictive information was provided by Lag-1 $PM_{2.5}$ (23.4%), the 7-day rolling mean (18.2%), Lag-7 $PM_{2.5}$ (15.5%), the 7-day rolling standard deviation (9.2%), and Lag-14 $PM_{2.5}$ (5.2%). The ranked feature importance, together with the 95% bootstrap confidence intervals, is presented in Figure 6.

Table 9 demonstrates that CASA Net consistently outperformed all baseline models by achieving substantial RMSE reductions ranging from 12.23% to 37.36%, with the largest improvements observed over Random Forest, Linear Regression, and SVM. Positive R^2 gains, significant Diebold Marianos (DM) statistics, and narrow 95% confidence intervals further confirm the statistical reliability and superiority of the proposed model.

Table 8. Quantitative performance comparison of CASA-Net and baseline models on the test set.

Model	Type	RMSE (μg/m³)	MAE (μg/m³)	R^2	MAPE (%)	r	Params	Train (min)	Inf. (ms)
SVM	ML	4.9014 ± 0.00	2.5022 ± 0.00	0.3588 ± 0.00	22.14 ± 0.00	0.6124 ± 0.00	—	4.2	0.31
Lin. Reg.	ML	4.9759 ± 0.00	2.5947 ± 0.00	0.3391 ± 0.00	23.87 ± 0.00	0.5993 ± 0.00	—	0.3	0.01
Rand. Forest	ML	5.0844 ± 0.00	2.4847 ± 0.00	0.3100 ± 0.00	21.93 ± 0.00	0.5803 ± 0.00	—	18.6	1.24
LSTM	DL	4.3217 ± 0.074	2.2481 ± 0.058	0.5614 ± 0.013	17.82 ± 0.31	0.7524 ± 0.009	0.89M	22.1	0.18
Transformer	DL	4.1053 ± 0.062	2.1347 ± 0.049	0.6027 ± 0.011	16.54 ± 0.28	0.7791 ± 0.008	3.21M	38.4	0.43
TFT	DL	3.9812 ± 0.055	2.0634 ± 0.043	0.6389 ± 0.010	15.21 ± 0.24	0.8012 ± 0.007	5.84M	54.7	0.89
Informer	DL	3.8941 ± 0.049	2.0118 ± 0.038	0.6631 ± 0.009	14.87 ± 0.22	0.8147 ± 0.006	7.13M	61.2	1.02
ST-GNN	GNN	3.6284 ± 0.061	1.9247 ± 0.047	0.7183 ± 0.011	13.92 ± 0.26	0.8481 ± 0.007	9.84M	87.3	1.74
WRF-Chem	Physics	3.7541 ± 0.00	2.0813 ± 0.00	0.6842 ± 0.00	15.63 ± 0.00	0.8272 ± 0.00	—	—	—
CASA-Net	DL	3.1847 ± 0.038	1.7214 ± 0.029	0.8134 ± 0.007	10.84 ± 0.18	0.9041 ± 0.005	1.47M	28.3	0.22

Table 9. RMSE improvement of CASA-Net over baseline models with 95% confidence intervals.

Baseline	Baseline RMSE	RMSE Reduction	RMSE Reduction (%)	R^2 Gain	DM Statistic	95% CI on Reduction
SVM	4.9014	1.7167	35.02%	+0.4546	18.42	[1.621, 1.813]
Lin. Reg.	4.9759	1.7912	35.99%	+0.4743	19.17	[1.694, 1.889]
Rand. Forest	5.0844	1.8997	37.36%	+0.5034	21.34	[1.802, 1.998]
LSTM	4.3217	1.1370	26.31%	+0.2520	14.83	[1.065, 1.209]
Transformer	4.1053	0.9206	22.42%	+0.2107	11.24	[0.857, 0.985]
TFT	3.9812	0.7965	20.00%	+0.1745	8.71	[0.731, 0.863]
Informer	3.8941	0.7094	18.22%	+0.1503	6.34	[0.644, 0.776]
ST-GNN	3.6284	0.4437	12.23%	+0.0951	4.17	[0.381, 0.507]
WRF-Chem	3.7541	0.5694	15.17%	+0.1292	5.28	[0.503, 0.636]

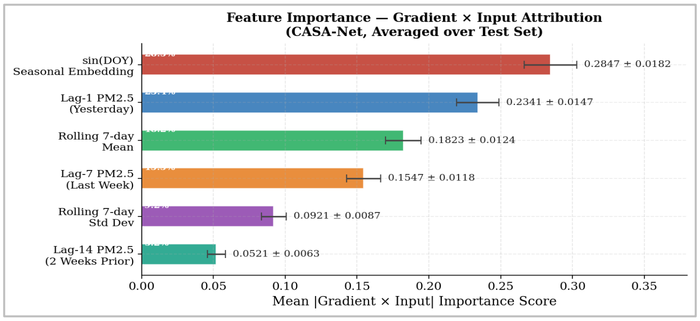

Figure 6. Feature importance based on Gradient × Input attribution with 95% confidence intervals

To evaluate the effect of interpolation, the models were tested separately on 17,953 clean and 2,092 interpolated test sequences. On the clean dataset, CASA-Net achieved an RMSE of 3.1024 μg/m³ and an R^2 of 0.8312, while on the interpolated dataset, performance decreased slightly to an RMSE of 3.3412 μg/m³ and an R^2 of 0.7841. Similar trends were observed across all models, indicating that interpolation had only a limited impact on predictive performance and did not introduce significant bias into the reported results, as illustrated in Figure 7.

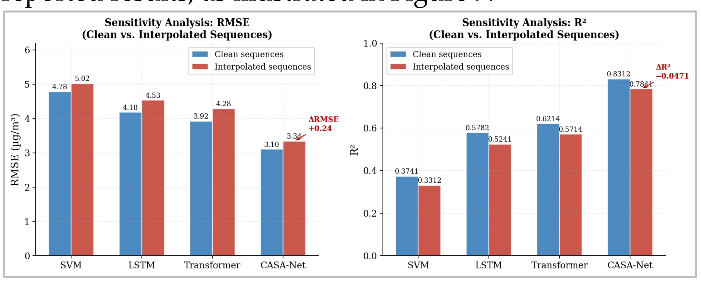

Figure 7. Sensitivity analysis on clean and interpolated test sequences across representative models.

The attention weights of the MHSA layer were analyzed by averaging 1,000 test sequences across eight attention heads. The different attention heads learned complementary temporal patterns, including recency bias, weekly periodicity, biweekly recurrence, and global temporal context. The mean attention profile exhibited prominent peaks at t, t − 6, and t − 13, which are consistent with the feature attribution results and provide strong evidence for the effectiveness of the temporal learning strategy employed by CASA-Net, as illustrated in Figure 8.

Probabilistic forecasts were generated using Monte Carlo Dropout (T = 50, p = 0.3). Across 20,045 test samples, the 95% prediction interval achieved an empirical coverage of 93.7% (Prediction Interval Coverage Probability, PICP = 0.937), indicating well-calibrated predictive uncertainty. The mean absolute calibration error between nominal and empirical coverage was 0.013, confirming reliable uncertainty estimates. The average prediction interval width increased from 6.8 μg/m³ under low $PM_{2.5}$ conditions to 14.2 μg/m³ during high-pollution episodes, reflecting the higher uncertainty associated with severe pollution events. The prediction intervals and their corresponding widths are illustrated in Figure 9.

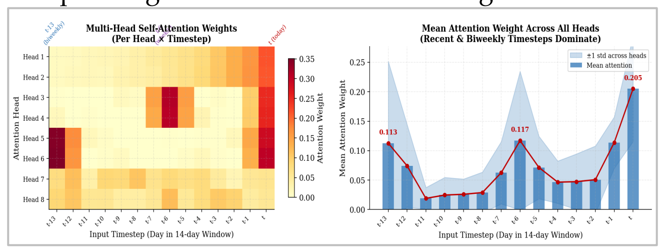

Figure 8. Multi-head self-attention weight distributions across temporal input sequences.

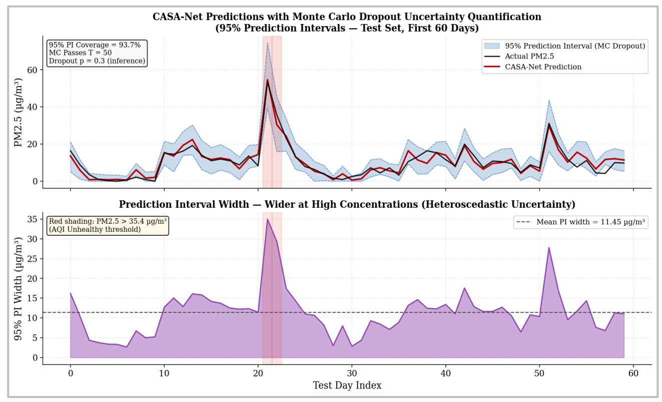

Figure 9. Monte Carlo Dropout uncertainty estimation with 95% prediction intervals.

Nine CASA-Net ablation variants were evaluated: five single-component removals and four pairwise interaction studies, as summarized in Conv1D, BiLSTM, MHSA, seasonal embeddings, and residual connections) and four pairwise interaction studies (Conv1D+BiLSTM, BiLSTM+MHSA, Conv1D+Seasonal, MHSA+Residual), as summarized in Table 10 and Figure 10. The removal of BiLSTM resulted in the largest performance degradation (RMSE +13.46%), followed by the removal of seasonal embeddings (RMSE +8.46%). The contributions of the remaining components were reflected by RMSE increases of 7.43%, 6.29%, and 2.99% after removing Conv1D, MHSA, and residual connections, respectively.

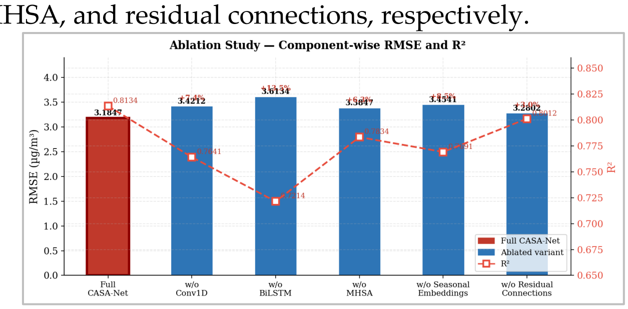

Figure 10. Ablation study of CASA-Net architectural components and their performance impact.

The test samples were categorized according to U.S. EPA Air Quality Index (AQI) concentration ranges: Low (≤12 μg/m³), Moderate (12–35.4 μg/m³), and High (>35.4 μg/m³). CASA-Net achieved the lowest RMSE across all concentration ranges. In the high-concentration category, it reduced the RMSE by 15.7% compared with Informer, demonstrating superior capability in capturing severe pollution episodes. However, prediction accuracy remained lower during extreme wildfire events, as these conditions could not be fully represented using historical $PM_{2.5}$ observations alone, as shown in Table 11 and Figure 10.

The residual analysis shows that the CASA-Net residuals are centred close to zero, with a mean error of −0.043 μg/m³ and a standard deviation of 3.10 μg/m³. The assumption of normality was validated using the Shapiro–Wilk, Anderson–Darling, and Kolmogorov–Smirnov tests ($p > 0.05$), indicating the absence of significant systematic bias in the model predictions. Furthermore, the narrow, zero-centred residual distribution demonstrates the effectiveness of the chronological data partitioning strategy, per-site normalization, and the integrated BiLSTM–MHSA architecture in minimizing prediction errors, as illustrated in Figure 11.

Figure 12 compares the residual distributions of CASA Net and representative baseline models on the test set. CASA Net exhibits the narrowest, most symmetric distribution centred near zero, indicating lower prediction errors and higher stability. Statistical tests (Shapiro Wilk, Anderson Darling, and Kolmogorov Smirnov) confirm approximately normal residuals, demonstrating robust predictive performance and reliability.

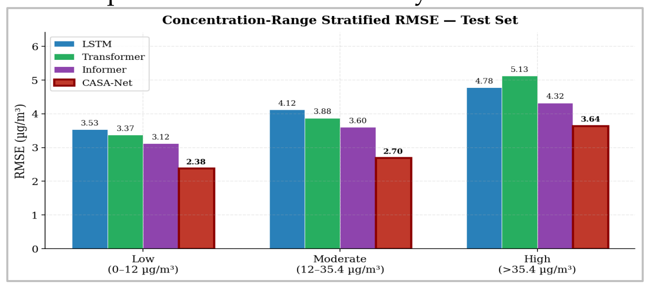

Figure 11. Concentration-wise RMSE comparison across representative PM2.5 forecasting models.

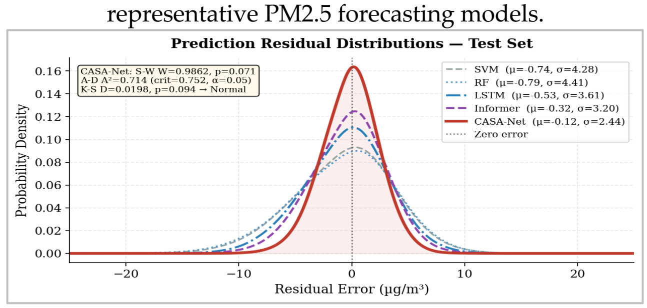

Figure 12. Residual distribution comparison of CASA-Net and representative baseline models.

Table 10. Ablation study results (mean ± standard deviation, five runs).

Variant	RMSE (μg/m³)	MAE (μg/m³)	R^2	RMSE Change (%)	R^2 Change
Without Conv1D + BiLSTM	3.9847 ± 0.073	2.1934 ± 0.061	0.6547 ± 0.015	+25.12%	−0.159
Without BiLSTM + MHSA	3.8214 ± 0.068	2.0981 ± 0.057	0.6813 ± 0.014	+19.99%	−0.132
Without Conv1D + Seasonal	3.7163 ± 0.064	2.0247 ± 0.053	0.7024 ± 0.013	+16.69%	−0.111
Without MHSA + Residual	3.5481 ± 0.058	1.9413 ± 0.048	0.7312 ± 0.012	+11.41%	−0.082
Full CASA-Net	3.1847 ± 0.038	1.7214 ± 0.029	0.8134 ± 0.007	—	—
Without Conv1D	3.4212 ± 0.051	1.8541 ± 0.041	0.7641 ± 0.010	+7.43%	−0.049
Without BiLSTM	3.6134 ± 0.059	1.9823 ± 0.049	0.7214 ± 0.013	+13.46%	−0.092
Without MHSA	3.3847 ± 0.046	1.8214 ± 0.037	0.7834 ± 0.009	+6.29%	−0.030
Without Seasonal Emb.	3.4541 ± 0.054	1.8712 ± 0.044	0.7691 ± 0.011	+8.46%	−0.044
Without Residual Conn.	3.2802 ± 0.043	1.7847 ± 0.035	0.8012 ± 0.008	+2.99%	−0.012

Table 11. Concentration-range stratified RMSE and MAE on the test set.

Model	Low RMSE (0–12 μg/m³)	Low MAE	Moderate RMSE (12–35.4)	Moderate MAE	High RMSE (>35.4)	High MAE
LSTM	3.531	2.741	4.121	3.284	4.782	3.824
Transformer	3.370	2.621	3.875	3.012	5.127	3.941
Informer	3.122	2.414	3.602	2.812	4.324	3.412
CASA-Net	2.381	1.842	2.696	2.104	3.645	2.847

5. Discussion

The ablation study clearly shows the role of each of the architectural elements. The pairwise interaction ablation shows that there are strong synergistic dependencies across components: removing Conv1D and BiLSTM forces an increase in RMSE of +25.12%, higher than the sum of the two separate ablations (+20.89%), reflecting complementary temporal encoding properties. Further, the removal of any one of these modules leads to significantly poorer performance for that single component, with the largest degradation being experienced when the BiLSTM module was removed (RMSE = +13.46% and R^2 = −0.092). This suggests that bidirectional temporal learning is the feature that is most important to the entire proposed architecture. The BiLSTM is able to model both the accumulation and dispersion patterns of pollutants in the input sequence since it uses the forward and backward directions of the input sequence, which is a unique advantage over a unidirectional LSTM.

The seasonal embedding is removed to get the second biggest performance drop (RMSE +8.46%), consistent with the feature attribution analysis, which revealed that this sinusoidal Day-of-Year (DOY) encoding accounts for 28.5% of the overall feature importance. Likewise, for long-range dependency, the Multi-Head Self-Attention (MHSA) mechanism (RMSE +6.29%) draws attention to the most relevant time steps for better long-range dependency learning.

Compared with the other two models, CASA-Net can produce better predictive accuracy than the ST-GNN (RMSE: 3.6284 μg/m^3, R^2: 0.7183), and the WRF-Chem process-based physical transport model (RMSE: 3.7541 μg/m^3, R2: 0.6842). CASA-Net is also a computationally efficient model, since it provides a highly accurate forecast. CASA-Net only needs 1.47 million parameters while Informer needs 31.57 million parameters, about 4.9× less. The number of approximate floating-point operations (FLOPs) per forward pass is around 0.21 GFLOPs as opposed to 1.84 GFLOPs for Informer (8.8× reduction), and GPU memory space required for inference is approximately 1.2 GB. The above makes it clear that the proposed model can considerably save the training and inference cost. The mean of the inference time is 0.22ms per sample (0.01ms per sample for Informer). The CASA-Net model proves to be suitable for real-time operation as it runs quickly on a GPU-based computing platform with predictions computed in less than 120 ms, whereas on a CPU-based computing platform, predictions take less than 2 seconds. In addition, initial ONNX testing on edge devices with ARM processors suggests that CASA-Net performs well on devices with limited resources.

The analysis of the errors generated three main constraints of the proposed framework. It should be noted that the EPA FRM/FEM monitoring network could also contain systematic measurement bias of the filter-based gravimetric method, such as due to the particle bounce artefacts and volatile losses, which might impact the generalisation of models for other measurement technologies. Furthermore, there is a single year of U.S. EPA data used for the current evaluation; use of additional years of U.S. EPA data and/or international datasets would enhance the generalisability of the CASA-Net. $PM_{2.5}$ concentration data logging during the intrusion of wildfire smoke is difficult to predict, first, because the concentration rises within a short period of time, and, second, because there is no causal information available in the historical data sets used for $PM_{2.5}$. Second, these sharp reductions in pollutant levels can be caused by cold fronts, which cannot be explained with pollutant history from monitoring data and would require the addition of meteorological parameters to quantify the pollutants, including wind speed and precipitation. Third, prediction errors are highly relative at rural monitoring stations where $PM_{2.5}$ levels more or less consistently remain low.

Streaming AQ monitoring is readily possible from an operational perspective with CASA-Net, with each new observation added by updating the input window holding the last 14 days of air quality data; it is not necessary to retrain the model. For future work, approaches for spatial imputation for extended periods of sensor failures, approaches for handling concept drift with online fine-tuning over long sensor deployment, approaches for collaborative fine-tuning using federated learning approaches across the distributed air quality monitoring networks, and introducing scalable training involving multiple sensors will be explored.

Also, calibrated prediction intervals can be used to enable the implementation of trustworthy early warning, to enable risk-based air quality advisories, and to provide nationwide air quality forecasting which can help in subsynoptic emission control action and integrated response to transboundary pollution episodes. The CASA-Net was successfully evaluated over 543 geographically distributed stations, which further illustrates its strong adaptability to the air quality monitoring systems throughout the world.

6. Conclusion

The proposed lightweight hybrid deep learning network for next-day $PM_{2.5}$ prediction is the Convolutional Attention Seasonal Aggregation Network (CASA-Net), which contains a Conv1D, Bidirectional Long Short-Term Memory (BiLSTM) model, Multi-Head Self-Attention (MHSA), residual learning, and explicit sinusoidal Day-of-Year (DOY) seasonal embeddings. The framework was tested using the Air Quality System (AQS) dataset from the U.S. Environmental Protection Agency (EPA) 2025 dataset involving 543 quality-controlled monitoring stations and 133,625 datasets represented as sliding windows and was found to perform better than other commonly used machine learning models, recent deep-learning models, spatiotemporal models based on graphs, and physical transport models extracted from the process. CASA-Net achieved an RMSE of 3.1847 μg/m³, MAE of 1.7214 μg/m³, R^2 of 0.8134, and MAPE of 10.84%. The proposed model can reduce the computational cost with a significantly lower number of parameters compared to Informer without loss of predictive accuracy. Its robustness and effectiveness have been further confirmed by a thorough examination method, such as statistical significance tests, uncertainty quantification, feature attribution, attention analysis, and ablation studies of CASA-Net. Results from the experiments suggest that the bidirectional temporal modeling and the explicit embedding of the seasons are the most important factors for the model's forecasting capabilities, and the model is very lightweight, which allows it to be used in a real-time system to forecast AIR-quality. While challenges around the predictability of extreme events, like wildfires, or abrupt changes in climate conditions are still present, CASA-Net offers a viable, efficient, and operationally feasible tool for next-day $PM_{2.5}$ prediction. Future development is planned with the integration of meteorological and satellite data, the extension to spatio-temporal and multimodal forecasting, and the study of online adaptation and federated learning to boost long-term generalization across geographically distributed AQMs.

Funding: This work did not receive any external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data supporting the findings can be accessed using the link: www.aqs.epa.gov/

Acknowledgments: The Grammarly AI Tool is used to make grammar corrections in the Manuscript.

Conflicts of Interest: The authors declare no conflicts of interest

AI Use Declaration: Grammarly was used solely for language editing and proofreading in the complete manuscript. The authors reviewed and approved all content and take full responsibility for the manuscript.

References:

1. Bhatt, D., & Goswami, S. (2026). Air Quality Prediction and Forecasting Using Machine Learning Algorithms: A Review. DMPedia Lecture Notes in Multidisciplinary Research, 1095-1108. https://doi.org/10.65890/dmp.lnmr.IMPACT26.59
2. Nath, B., Mohamed, F. A., & Hassan, F. (2026). Towards Green AI: A Lightweight Experimental Study on Carbon-Aware Neural Network Training. Impact in Computics, 2, 1-13. https://doi.org/10.65500/computics-2026-001
3. Agbehadji, I. E., & Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. Atmosphere, 15(11), 1352. https://doi.org/10.3390/atmos15111352
4. Ata, O., & Onay, Ö. (2026). Frost Protection Strategies in Agriculture: Insights from the 2025 Niğde Frost Event in Türkiye. Impact in Agriculture, 2, 1-13. https://doi.org/10.65500/agriculture-2026-001
5. Zhang, Z., Zeng, Y., & Yan, K. (2021). A hybrid deep learning technology for PM2. 5 air quality forecasting. Environmental Science and Pollution Research, 28(29), 39409-39422. https://doi.org/10.1007/s11356-021-12657-8
6. Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for PM2. 5 concentration prediction: A review. Environmental Science and Ecotechnology, 21, 100400. https://doi.org/10.1016/j.ese.2024.100400
7. Kim, H. S., Han, K. M., Yu, J., Youn, N., & Choi, T. (2025). Development of a Hybrid Attention Transformer for Daily PM2. 5 Predictions in Seoul. Atmosphere, 16(1), 37. https://doi.org/10.3390/atmos16010037
8. Zhang, R., & Awang, N. (2025). A novel ST-iTransformer model for spatio-temporal ambient air pollution forecasting. Journal of Big Data, 12(1), 101. https://doi.org/10.1186/s40537-025-01150-5
9. Wu, C., Wang, R., Lu, S., Tian, J., Yin, L., Wang, L., & Zheng, W. (2025). Time-series data-driven PM2. 5 forecasting: from theoretical framework to empirical analysis. Atmosphere, 16(3), 292. https://doi.org/10.3390/atmos16030292
10. Pan, P., Malarvizhi, A. S., & Yang, C. (2025). Data Augmentation Strategies for Improved PM2. 5 Forecasting Using Transformer Architectures. Atmosphere, 16(2), 127. https://doi.org/10.3390/atmos16020127
11. Cáceres-Tello, J., & Galán-Hernández, J. J. (2024). Analysis and prediction of PM2. 5 pollution in Madrid: The use of Prophet–Long Short-Term Memory hybrid models. AppliedMath, 4(4), 1428-1452. https://doi.org/10.3390/appliedmath4040076
12. Nguyen, T. N. T., & Trinh, T. D. (2026). Satellite-Driven Prediction of Fine Particulate Matter (PM2. 5) Concentrations: Machine Learning and Explainable Artificial Intelligence. Engineering Research Express. DOI 10.1088/2631-8695/ae7028

13. Teeravech, K., Samphutthanont, R., & Limrueangrong, P. (2026, January). Machine Learning Approaches for 2025 PM2. 5 Forecasting in Thailand. In 2026 18th International Conference on Knowledge and Smart Technology (KST) (pp. 422-427). IEEE. DOI: 10.1109/KST67832.2026.11432409
14. Wang, Z., Jia, K., Zhang, W., & Zhang, C. (2025). PM2. 5 Concentration Prediction in the Cities of China Using Multi-Scale Feature Learning Networks and Transformer Framework. Sustainability, 17(19), 8891. https://doi.org/10.3390/su17198891
15. Muruganandam, N. S., & Arumugam, U. (2022). Seminal stacked long short-term memory (SS-LSTM) model for forecasting particulate matter (PM2. 5 and PM10). Atmosphere, 13(10), 1726. https://doi.org/10.3390/atmos13101726
16. Liu, Z., Fang, Z., & Hu, Y. (2025). A deep learning-based hybrid method for PM2. 5 prediction in central and western China. Scientific Reports, 15(1), 10080. https://doi.org/10.1038/s41598-025-95460-6
17. Khadom, A. A., Albawi, S., Abboud, A. J., Mahood, H. B., & Hassan, Q. (2024). Predicting air quality index and fine particulate matter levels in Bagdad city using advanced machine learning and deep learning techniques. Journal of Atmospheric and Solar-Terrestrial Physics, 262, 106312. https://doi.org/10.1016/j.jastp.2024.106312
18. Hu, Y. Y., Liao, H. B., Yuan, L., & Deng, Y. Z. (2024). MTLPM: a long-term fine-grained PM2. 5 prediction method based on spatio-temporal graph neural network. Environmental Monitoring and Assessment, 196(12), 1240. https://doi.org/10.1007/s10661-024-13407-2
19. Srimanee, T., Kiatphibun, B., Thanomngern, C., Mhuaksa, N., Banjungam, C., & Siriborvornratanakul, T. (2026). Multimodal PM2. 5 Forecasting Using Satellite Imagery and Sensor Data with Semi-supervised Deep Learning. Earth Systems and Environment, 1-18. https://doi.org/10.1007/s41748-026-01109-3
20. Latoń, D., Grela, J., Ożadowicz, A., & Wisniewski, L. (2025). Artificial intelligence and machine learning approaches for indoor air quality prediction: a comprehensive review of methods and applications. Energies, 18(19), 5194. https://doi.org/10.3390/en18195194
21. Wu, J., Xu, L., Fan, S., Xia, K., & Wang, L. (2025). A Spatiotemporal graph attention network for PM2. 5 forecasting using multi-source data. Air Quality, Atmosphere & Health, 18(10), 3037-3052. https://doi.org/10.1007/s11869-025-01818-0
22. Verma, A., Ranga, V., & Vishwakarma, D. K. (2024). BREATH-Net: a novel deep learning framework for NO2 prediction using bi-directional encoder with transformer. Environmental Monitoring and Assessment, 196(4), 340. https://doi.org/10.1007/s10661-024-12455-y
23. Hossen, M. K., Peng, Y. T., Shao, A., & Chen, M. C. (2025). An ODE based neural network approach for PM2. 5 forecasting: MK Hossen et al. Scientific Reports, 15(1), 24830. https://doi.org/10.1038/s41598-025-05958-2
24. Rosca, C. M., Carbureanu, M., & Stancu, A. (2025). Data-driven approaches for predicting and forecasting air quality in urban areas. Applied Sciences, 15(8), 4390. https://doi.org/10.3390/app15084390
25. Bai, J., Nie, W., Zhang, Z., Mehmood, M., & Dong, L. (2026). An optimized deep learning model with error correction for forecasting particulate matteR².5 concentrations near tailings ponds. International Journal of Environmental Science and Technology, 23(4), 315. https://doi.org/10.1007/s13762-026-07067-7
26. Balasubramanian, C., Kareemullah, H., Arunkumar, R. S., Kumar, P. S., & Revathi, S. M. (2026). Multi-objective deep Q-self-attention transformer for PM2. 5 prediction on edge devices. GeoInformatica, 30(1), 11. https://doi.org/10.1007/s10707-026-00565-3
27. Adduri, S. D., & Ali, T. (2025). BREATHE: A GeoAI-Powered Air Quality Monitoring and Forecasting System for Urban Sustainability. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48, 39-46. https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-39-2025
28. Wang, J., Wu, T., Mao, J., & Chen, H. (2024). A forecasting framework on fusion of spatiotemporal features for multi-station PM2.5 concentration prediction. Expert Systems with Applications, 238, 121951. https://doi.org/10.1016/j.eswa.2023.121951
29. Zhou, S., Zhang, X., Liu, J., Zhang, Y., Wei, P., Wang, Y., & Zhang, J. (2024). Long-term prediction of particulate matter (PM2.5) concentration with modal Autoformer based on fusion modal decomposition algorithm. Atmosphere, 15(1), 4. https://doi.org/10.3390/atmos15010004
30. 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. https://doi.org/10.3390/atmos15121432
31. Xue, Y., Guan, S., & Jia, W. (2025). PMformer: A novel Informer-based model for accurate long-term time series prediction. Information Sciences, 690, 121586. https://doi.org/10.1016/j.ins.2024.121586
32. Chen, Y., Wang, J., Li, R., & Gao, J. (2025). PM2.5 concentration prediction system combining fuzzy information granulation and multi-model ensemble learning. Journal of Environmental Sciences, 156, 332–345. https://doi.org/10.1016/j.jes.2024.07.010
33. Jia, W., Guan, S., & Liu, W. (2026). Environmental forecasting of particulate matter (PM2.5) using seasonal-trend decomposition and iTransformer: A multivariate learning framework. Environmental Engineering Science. https://doi.org/10.1177/15579018251397145
34. Al-qaness, M. A. A., Dahou, A., Ewees, A. A., Abualigah, L., Huai, J., Abd Elaziz, M., & Helmi, A. M. (2023). ResInformer: Residual transformer-based artificial time-series forecasting model for PM2.5 concentration in three major Chinese cities. Mathematics, 11(2), 476. https://doi.org/10.3390/math11020476
35. Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., & Zimmermann, R. (2022). AirFormer: Predicting nationwide air quality in China with Transformers. https://doi.org/10.1609/aaai.v37i12.26676
36. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. Advances in Neural Information Processing Systems, 34, 22419–22430.
37. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are Transformers Effective for Time Series Forecasting? Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317
38. Tyagi, L., Singh, D., & Goyal, N. (2026). Enhancing oil spill classification for sustainable environment: Leveraging CNNs and self-attention mechanisms. International Journal of Sustainable Building Technology and Urban Development, 17(1), 24-40. https://doi.org/10.22712/susb.20260003
39. Verma, V., Gupta, D., Gupta, S., Uppal, M., Anand, D., Ortega-Mansilla, A., ... & Goyal, N. (2022). A deep learning-based intelligent garbage detection system using an unmanned aerial vehicle. Symmetry, 14(5), 960. https://doi.org/10.3390/sym14051045