Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hybrid Physics-Informed Machine Learning Framework for Interpretable Ozone Formation Modelling in a Coastal Urban Location

Ebin Antony
Kannur University
Sunil Kumar RK
Kannur University
Nishanth T
nisthu.t@gmail.com

Sree Krishna College Guruvayur

Research Article

Keywords: Surface Ozone, Air Quality Modeling, Physics-Informed Neural Networks, Symbolic Regression, Atmospheric Chemistry, Environmental Time-Series Analysis

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9351951/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hybrid Physics-Informed Machine Learning Framework for Interpretable Ozone Formation Modelling in a Coastal Urban Location

Abstract:

Surface ozone (O_3) formation is governed by highly nonlinear interactions among precursor emissions, meteorological conditions, and photochemical processes, making accurate prediction and interpretation challenging. This study presents a hybrid data-driven framework that integrates machine learning (XGBoost and Long Short-Term Memory networks), Physics-Informed Neural Networks (PINNs), and Symbolic Regression (SR) to model ozone dynamics over a coastal urban environment in Kannur in south India, using high-resolution (15-minute) observations. Among the models, XGBoost achieved the highest predictive accuracy (R^2 = 0.98; RMSE = 0.60 ppbv), effectively capturing nonlinear relationships between ozone and its precursors. The data-driven PINN demonstrated comparable accuracy (R^2 = 0.98; RMSE = 0.89 ppbv) while ensuring physical consistency by embedding photochemical constraints within the learning process. Symbolic regression further provided interpretable analytical expressions that revealed key ozone production and loss pathways, consistent with established atmospheric chemistry, including NO_x titration, photolysis, and aerosol-mediated suppression. Feature attribution using SHAP identified solar radiation, temperature, and NO_2 as dominant drivers of ozone formation, while particulate matter exhibited a moderating effect through radiative attenuation. The results highlight that hybrid physics-guided machine learning approaches can simultaneously achieve high predictive accuracy, physical realism, and chemical interpretability. This study demonstrates that integrating data-driven models with physical constraints and equation discovery offers a robust and scalable framework for atmospheric chemistry modeling, with significant implications for air quality forecasting, environmental policy, and climate–air quality interactions.

Keywords:

Surface Ozone; Air Quality Modeling; Physics-Informed Neural Networks; Symbolic Regression; Atmospheric Chemistry; Environmental Time-Series Analysis

1. Introduction

Tropospheric ozone (O_3) is a secondary pollutant formed through photochemical reactions involving nitrogen oxides (NO_x), volatile organic compounds (VOCs), carbon monoxide (CO), and other precursors in the presence of solar radiation. Unlike primary pollutants that are directly emitted, ozone arises from nonlinear atmospheric chemistry, making its prediction and regulation particularly challenging. As a highly reactive oxidant, ozone poses significant risks to human health, contributing to respiratory diseases and premature mortality (Jerrett et al., 2009). It also adversely affects agricultural productivity by reducing crop yields (Avnery et al., 2011) and contributes to climate forcing as a short-lived greenhouse gas (Stevenson et al., 2006).

Particulate matter (PM), particularly $PM_{2.5}$ and PM_{10}, plays a dual role in atmospheric processes. In addition to its direct health impacts, PM influences ozone chemistry by modifying actinic flux through scattering and absorption of radiation and by facilitating heterogeneous reactions on particle surfaces that can suppress photolysis rates (Dickerson et al., 1997). This complex coupling between ozone and particulate matter highlights the importance of understanding their interactions for both scientific research and policy development.

The chemical basis of ozone formation is described by the Leighton relationship (Leighton, 1961), which links ozone (O_3), nitrogen monoxide (NO), and nitrogen dioxide (NO_2) under photostationary equilibrium. Subsequent developments, such as the Carbon Bond Mechanism (Whitten et al., 1980), expanded the representation of VOC oxidation pathways. These frameworks demonstrate that ozone formation occurs under two distinct regimes: NO_x-limited conditions, where reducing NO_x emissions decreases ozone concentrations, and VOC-limited conditions, where NO_x reductions may paradoxically increase ozone levels (Sillman, 1999). Accurately capturing this nonlinear behavior remains a major challenge, particularly under conditions of high aerosol loading, which can further suppress ozone production (Tie et al., 2005).

To simulate these processes, Chemical Transport Models (CTMs) such as the Community Multiscale Air Quality (CMAQ) model (Byun & Schere, 2006) and WRF-Chem (Grell et al., 2005) are widely used. Although CTMs are physically based and capable of coupling atmospheric chemistry with meteorology, they are computationally intensive and sensitive to uncertainties in emission inventories, boundary conditions, and aerosol–radiation interactions. Furthermore, their ability to accurately represent PM–ozone feedbacks is still limited, often leading to discrepancies between modeled and observed ozone concentrations, especially in urban and industrial regions.

Recent advances in machine learning (ML) have introduced new opportunities for atmospheric chemistry modeling. Data-driven approaches can capture complex nonlinear relationships directly from observations, thereby reducing reliance on uncertain mechanistic representations. Studies have shown that neural networks, ensemble tree methods, and hybrid ML techniques can achieve high accuracy in short-term ozone forecasting (Zhan et al., 2018; Nowack et al., 2018). However, many of these models function as "black boxes," offering limited interpretability despite their predictive capability (McGovern et al., 2019). This lack of interpretability poses a significant limitation for policy applications, where understanding chemical mechanisms and precursor contributions is essential for designing effective emission control strategies.

To address these challenges, this study proposes a hybrid modeling framework that integrates predictive accuracy with chemical interpretability. Specifically, we combine tree-based ensemble methods (XGBoost) and recurrent neural networks (LSTM) with explainability tools such as SHapley Additive exPlanations (SHAP), symbolic regression for empirical rate law discovery, and Physics-Informed Neural Networks (PINNs), which incorporate known physical constraints into the learning process. This integrated approach not only enhances predictive performance but also enables the extraction of chemically meaningful relationships consistent with, or extending beyond, traditional kinetic models. Thus, machine learning is positioned not merely as a forecasting tool but as a means to advance atmospheric chemistry understanding and support evidence-based air quality management.

2. Recent developments in Ozone forecasting using machine learning techniques

Ground-level ozone (O_3) is a secondary pollutant produced through photochemical reactions involving nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under solar radiation. The theoretical understanding of ozone formation is rooted in atmospheric photochemistry, which is governed by complex nonlinear differential equations describing chemical kinetics and transport processes (Seinfeld & Pandis, 2016). Traditional modeling approaches, including the Community Multiscale Air Quality (CMAQ) model and the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem), have been extensively used to simulate ozone concentrations under varying meteorological conditions (Byun & Schere, 2006). However, these deterministic models require detailed emission inventories, accurate chemical mechanisms, and substantial computational resources, limiting their applicability for real-time forecasting and regional adaptation (Jacob, 1999).

These limitations have driven the development of data-driven and hybrid approaches that integrate empirical observations with mechanistic insights to improve both efficiency and

interpretability (Zhang et al., 2018). In recent years, machine learning (ML) techniques have been increasingly applied to model ozone dynamics. Ensemble methods such as Random Forest and Extreme Gradient Boosting (XGBoost) have demonstrated strong predictive performance due to their ability to capture nonlinear relationships between pollutants and meteorological variables (Fan et al., 2023). Comparative studies indicate that XGBoost often outperforms traditional statistical approaches, including multiple linear regression and support vector machines (Liu et al., 2020).

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, are effective in capturing temporal dependencies in ozone time-series data (Hu et al., 2021; Antony, E., 2025). These models can learn diurnal and seasonal variations directly from data. However, purely data-driven approaches are prone to overfitting and may violate known physical and chemical principles, resulting in physically inconsistent predictions (Kumar et al., 2022). This has led to growing interest in physics-guided machine learning approaches that incorporate domain knowledge into model development (Raissi et al., 2019).

Interpretability has emerged as a critical requirement in ML-based air quality modeling, particularly for policy applications. Explainable Artificial Intelligence (XAI) methods, such as SHapley Additive exPlanations (SHAP), are widely used to quantify the contributions of key variables-including NO_2, solar radiation, temperature, and $PM_{2.5}$ to ozone concentrations (Lundberg & Lee, 2017). These methods provide both local and global interpretability, enabling the identification of key drivers of ozone variability. However, SHAP explains statistical associations rather than causal relationships and may produce biased attributions in the presence of correlated features (e.g., temperature and solar radiation) (Molnar, 2022). Therefore, careful feature selection and pre-processing remain essential.

Symbolic Regression (SR), often implemented using genetic programming, has emerged as a promising technique for deriving explicit analytical expressions that describe pollutant behavior. SR can uncover interpretable relationships, such as empirical reaction-rate equations linking NO_2, VOCs, solar radiation, and O_3 (Schmidt & Lipson, 2009). For example, Alam et al. (2022) demonstrated that SR can identify empirical rate laws consistent with established chemical kinetics. In air quality studies, SR has also been used to develop reduced-form models that approximate the behavior of complex CTMs while maintaining interpretability (Li & Wang, 2022).

Physics-Informed Neural Networks (PINNs) incorporate governing physical laws directly into neural network training by embedding differential equations into the loss function (Raissi, Perdikaris, & Karniadakis, 2019). This ensures that model predictions are consistent

with conservation laws and chemical kinetics, even when trained on limited data. This concept has been further generalized under Physics-Informed Machine Learning (Karniadakis et al., 2021). In atmospheric applications, PINNs have been used to model pollutant dispersion (Wang et al., 2022), estimate reaction rate constants (Luo et al., 2024), and enforce ozone–NO_x photochemical constraints (Antony, E. et al., 2025).

Hybrid modeling frameworks that integrate machine learning with physical models are increasingly recognized as a promising direction. These approaches often combine the predictive strength of XGBoost with the physical consistency of PINNs, achieving a balance between accuracy and interpretability (Chen et al., 2023). Comparative studies across regions such as Beijing–Tianjin–Hebei and Delhi–NCR indicate that hybrid models outperform both CTMs and traditional statistical approaches under varying emission and meteorological conditions (Zhang et al., 2022). More recently, Temporal Convolutional Networks (TCNs) and Transformer-based architectures have been explored as alternatives to LSTM models for ozone forecasting (Han et al., 2023). While these models improve the representation of long-range temporal dependencies, they generally lack inherent physical constraints, reinforcing the importance of physics-informed approaches such as PINNs. Despite significant progress, several challenges remain. Many machine learning models, although highly accurate, fail to maintain physical consistency with established photochemical principles. Interpretability also remains limited, as methods such as SHAP and permutation importance provide partial insights without yielding explicit chemical relationships. Additionally, parameter discovery—particularly the estimation of kinetic rate constants from observational data—remains underexplored.

Furthermore, fully integrated hybrid frameworks that combine feature importance analysis (e.g., XGBoost), temporal modeling (e.g., LSTM or TCN), symbolic regression for empirical law extraction, and physics-based constraints through PINNs are still scarce. To address these gaps, the present study proposes a comprehensive hybrid framework integrating XGBoost, SHAP, symbolic regression, and PINNs to model ozone formation. This approach aims to achieve a balance between predictive accuracy, interpretability, and physical consistency, thereby providing a robust and scientifically grounded understanding of ozone dynamics under diverse atmospheric conditions.

3. Methodology

This section describes the framework developed for modeling ozone (O_3) and PM_{10} formation using machine learning and data-driven chemical inference. The proposed hybrid framework integrates machine learning techniques with physics-based modeling to simulate ozone formation processes (Figure 1). Air quality and meteorological data collected from Kannur, Kerala, at 15-minute intervals are first subjected to preprocessing, including data cleaning, normalization, and feature selection. The XGBoost–SHAP module is employed to identify key predictors, such as NO_2, temperature, and solar radiation, thereby providing insights into the relative importance of input variables. Subsequently, a Long Short-Term Memory (LSTM) network is used to capture temporal dependencies and diurnal variations in pollutant concentrations. To ensure physical consistency, a Physics-Informed Neural Network (PINN) is implemented by embedding governing ozone photochemical kinetics into the model training process. The PINN is trained over 4000 epochs to achieve stable convergence while preserving adherence to known chemical constraints. In parallel, a symbolic regression module is applied to derive interpretable analytical expressions that describe the relationships between pollutants and meteorological parameters, enabling empirical chemical inference. Model performance is evaluated using standard statistical metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), to ensure robustness and reliability. The final framework provides both high predictive accuracy and physically interpretable insights, supporting informed decision-making in air quality management.

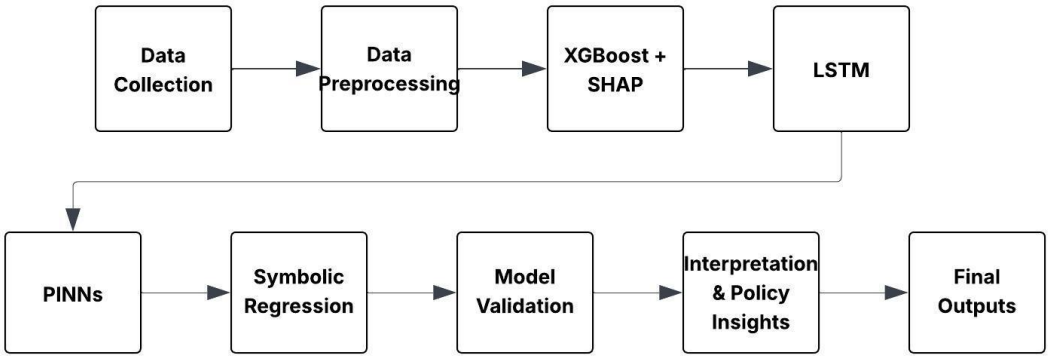


Figure 1: System Architecture of the Hybrid Ozone Modeling Framework

3.1 Dataset and Preprocessing

The dataset utilized in this study was obtained from an automated air quality monitoring station located in Kannur District in South India, representing a coastal urban environment influenced by both anthropogenic emissions and meteorological conditions. Measurements were recorded at 15-minute intervals, providing high temporal resolution for capturing short-term variability in atmospheric composition. The dataset comprises key air pollutants, including ozone (O_3), nitrogen oxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), ammonia (NH_3), and particulate matter ($PM_{2.5}$ and PM_{10}), along with relevant meteorological parameters such as temperature, relative humidity, and solar radiation. Let the dataset be denoted as:

$$D = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{N}$$

Where $x^{(i)} \in \mathbb{R}^d$ is the feature vector at time step i, consisting of pollutant concentrations (e.g., NO, NO_2, CO, $PM_{2.5}$, PM_{10}) and meteorological variables (e.g., temperature, solar radiation, relative humidity). $y^{(i)} \in \mathbb{R}$ is the target variable (Ozone), and N is the number of observations.

Missing values were imputed via linear interpolation across time, and all features were normalized using Min–Max scaling (Ebin Antony et al., 2021):

$$x_j^{(i)\prime} = \frac{x_j^{(i)} - \min(x_j)}{max}$$

Additionally, interaction terms such as the VOC/NO_x ratio were included to capture nonlinear precursor dynamics relevant to ozone chemistry.

3.2 Feature Importance via XGBoost and SHAP

3.2.1 XGBoost Model

The eXtreme Gradient Boosting (XGBoost) is a scalable ensemble learning algorithm that optimizes an additive collection of regression trees for high predictive performance (Chen & Guestrin, 2016). The model prediction is defined as:

$$\hat{y}^{(i)} = \sum_{k=1}^{K} f_k\left(x^{(i)}\right), \qquad f_k \in F,$$

where f_k represents a regression tree and $\mathcal{F}$ is the functional space of Classification and Regression Trees (CART). The learning objective is formulated as:

$$L(\phi) = \sum_{i=1}^{N} \ell\left(y^{(i)}, \hat{y}^{(i)}\right) + \sum_{k=1}^{K} \Omega(f_k),$$

with convex loss function ℓ (mean squared error) and regularization term:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2,$$

where T is the number of leaf nodes and w_j are the corresponding leaf weights. The regularization term controls model complexity and prevents overfitting, improving the model's generalization capability (Chen & Guestrin, 2016).

3.2.2 SHAP Attribution

To interpret feature contributions, SHapley Additive exPlanations (SHAP) values are computed to estimate the marginal impact of each input feature on the model output (Lundberg & Lee, 2017). The SHAP representation is given as:

$$\hat{y} = \emptyset_0 + \sum_{j=1}^{d} \emptyset_j x_j$$

where ϕ_j quantifies the contribution of feature x_j to the prediction. This framework enables both global and local interpretability by ranking pollutant and meteorological drivers of ozone formation according to their relative importance.

3.3 Time-Series Prediction using LSTM

To capture the temporal evolution of pollutant concentrations, Long Short-Term Memory (LSTM) networks are employed. LSTMs are a class of recurrent neural networks (RNNs) designed to overcome the vanishing gradient problem by introducing gated mechanisms that regulate information flow through time (Hochreiter & Schmidhuber, 1997). Given a windowed sequence of feature vectors $\{x^{(t-k)}, \ldots, x^{(t)}\}$, the model predicts future values $\hat{y}_{(t+1)}$ using internal memory cell states. The hidden states are computed as follows:

$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$ (Forget gate)

$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$ (Input gate)

$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$ (Candidate cell state)

$\tilde{c}_t = tanh(W_c x_t + U_c h_{t-1} + b_c)$ (Cell state update)

$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$ (Output gate)

$h_t = o_t \odot \tanh(c_t)$ (Hidden state)

Finally, the ozone concentration is predicted as: $\hat{y}_{(t+1)} = W_y h_t + b_y$

This formulation enables short-term forecasting of pollutant dynamics by learning temporal dependencies and preserving long-term memory of past atmospheric states (Graves, 2013; Hu et al., 2021).

3.4 Symbolic Regression for Reaction Rate Discovery

To recover interpretable kinetic-like equations, symbolic regression is applied to approximate the governing dynamics of ozone:

$$\frac{d[O_3]}{dt} = f(x),$$

where $f(x)$ is an expression drawn from the function space $\mathcal{F}$ of candidate mathematical forms (e.g., products, sums, exponentials). The optimization problem is expressed as:

$$\min_{f \in F} \sum_{i=1}^{N} \left(\frac{d[O_3]^{(i)}}{dt} - f(x^{(i)}) \right)^2.$$

A representative from derived in this study is:

$$\frac{d[O_3]}{dt} \approx f(x) = k_1[NO_2][SR]e^{-\beta_1[PM_{2.5}]} - k_2[NO][O_3] - k_3[PM_{2.5}][O_3] - k_4[CO][O_3] - k_5[SO_2] - k_6[NH_3] - k_7[PM_{10}] + k_8[Temp] + k_9[RH],$$

where each term corresponds to a known physical or chemical process—such as NO_2 photolysis, NO titration, heterogeneous losses, or temperature dependence. The coefficients $\{k_i\}$ are optimized using environmental data to minimize the residual error. Table 1 is the summary of ozone formation and loss mechanisms derived from symbolic regression, showing corresponding mathematical terms, formulas, and their underlying physical or chemical interpretations.

Table 1. Symbolic regression–derived ozone reaction terms and their physical or chemical interpretations.

Term	**Formula**	**Physical/Chemical Meaning**
O_3Production	$k_1[NO_2][SR]e^{-\beta_1[PM_{2.5}]}$	Photolysis of NO_2 (requires sunlight, attenuated by $PM_{2.5}$)
O_3 Loss (NO titration)	$-k_2[NO][O_3]$	NO reacts with O_3 to form NO_2 (ozone depletion)
O_3 Loss ($PM_{2.5}$)	$-k_3[PM_{2.5}][O_3]$	Heterogeneous reactions on $PM_{2.5}$ surfaces
O_3 Loss (CO oxidation)	$-k_4[CO][O_3]$	CO consumes O_3 in oxidation reactions
SO_2 Suppression	$-k_5[SO_2]$	SO_2 competes for oxidants, reducing O_3 formation

NH_3 Suppression	$-k_6[NH_3]$	NH_3 reacts with radicals, slowing O_3 production
PM_{10} Scattering	$-k_7[PM_{10}]$	PM_{10} scatters sunlight, reducing photolysis rates
Temperature Effect	$+k_8[Temp]$	Higher temps accelerate gas-phase reactions
Humidity Effect	$+k_9[RH]$	H_2O can enhance or suppress radical chemistry

Where NO, NO_2, O_3, PM, SO_2, NH_3, Temp, RH, SR are concentration levels. Each term in the equation corresponds to known atmospheric reactions and physical effects (e.g., photolysis, titration, heterogeneous reactions). The coefficients (k_1,…k_9) are optimized parameters learned from data to minimize the residual. Symbolic regression, often implemented using genetic programming, has proven highly effective for discovering analytical rate laws and empirical governing equations in physics and chemistry (Schmidt & Lipson, 2009; Cranmer et al., 2020; Alam et al., 2022). In atmospheric applications, Symbolic Regression has been employed to derive interpretable ozone and aerosol dynamics consistent with kinetic mechanisms (Li & Wang, 2022).

3.5 Physics-Informed Neural Networks (PINNs)

To ensure physical consistency in the ozone modeling process, we incorporate the governing photochemical dynamics into a Physics-Informed Neural Network (PINN) framework. Let $\hat{y}(t)$ denote the predicted ozone concentration from the neural network. The total training objective combines both data fidelity and physical law conformity through a composite loss function:

$$L_{\text{total}} = \alpha L_{\text{data}} + (1-\alpha)L_{\text{phys}},$$

where $\alpha \in [0,1]$ is a balancing coefficient between the empirical accuracy and physical consistency terms.

The components of the loss function are defined as follows:

i. Data Loss:

$$L_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2,$$

ii. Physics Loss (Residual of Governing Equation):

$$L_{\text{phys}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{d\hat{y}^{(i)}}{dt} - f(x^{(i)})\right)^2,$$

where $f(x)$ represents the symbolic or mechanistic ozone chemistry expression derived from prior regression analysis.

This hybrid loss formulation ensures that model predictions remain chemically plausible even under conditions of sparse or noisy data (Raissi et al., 2019; Karniadakis et al., 2021).

iii. PINN Formulation for Ozone Dynamics

The temporal evolution of ozone concentration can be described by a physics-based ordinary differential equation (ODE) of the form:

$$\frac{d[O_3]}{dt} = f(t, [O_3], [NO], [NO_2], [VOCs], I_{UV}, [PM]; p),$$

where $f(\cdot)$ denotes the production–loss function, $I_{UV}$is solar irradiance, and $p = \{k_1, k_2, \dots\}$ represents kinetic rate parameters.

iv. Neural Network Representation

The ozone concentration is approximated by a neural network:

$$\hat{y}(t;\theta) \approx [O_3](t),$$

where θdenotes the network weights and biases. The network takes as inputs time $t$and exogenous features $x(t) = ([NO], [NO_2], [VOCs], I_{UV}, [PM], \dots)$:

$$\hat{y}(t;\theta) = N(t, x(t); \theta).$$

Using automatic differentiation (AD), the time derivative of the predicted state is computed as:

$$\frac{d\hat{y}}{dt}(t;\theta) = \frac{\partial N}{\partial t}(t, x(t);\theta) + \sum_j \frac{\partial N}{\partial x_j}(t, x(t);\theta)\frac{dx_j}{dt}.$$

In typical PINN formulations where $x(t)$ is treated as static input, this simplifies to:

$$\frac{d\hat{y}}{dt}(t;\theta) = \frac{d}{dt}N(t, x(t);\theta).$$

v. Residual Definition

The residual of the governing ODE at time $t$is defined as:

$$R(t;\theta,p) = \frac{d\hat{y}}{dt}(t;\theta) - f(t, \hat{y}(t;\theta), x(t); p).$$

For coupled systems (e.g., ozone–PM interactions), $R(t)$ becomes vector-valued and is evaluated component-wise.

vi. Composite Loss Components

The PINN training minimizes the total loss composed of three parts:

1. Data loss (supervised term):

$$L_{\text{data}}(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d}(y^{(i)} - \hat{y}(t_i;\theta))^2$$

2. Physics loss (residual at collocation points):

$$L_{\text{phys}}(\theta, p) = \frac{1}{N_c}\sum_{j=1}^{N_c}[\frac{d\hat{y}}{dt}(t_j^{(c)};\theta) - f(t_j^{(c)}, \hat{y}(t_j^{(c)};\theta), x(t_j^{(c)}); p)]^2$$

3. Boundary or initial condition loss:

$$L_{\text{bc}}(\theta) = \frac{1}{N_{bc}}\sum_{k=1}^{N_{bc}} \| \hat{y}(t_k^{(bc)};\theta) - y_{bc}^{(k)} \|^2$$

The final PINN loss function is therefore:

$$L_{\text{PINN}} = L_{\text{data}} + \lambda_{\text{phys}}L_{\text{phys}} + \lambda_{\text{bc}}L_{\text{bc}},$$

where λ_{phys} and λ_{bc} are scaling parameters that balance the influence of physical and boundary constraints.

3.6 Model Evaluation

Model performance is evaluated using standard statistical metrics, including:

i. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2},$$

ii. Coefficient of Determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{N}(y^{(i)} - \bar{y})^2},$$

iii. Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y^{(i)} - \hat{y}^{(i)}|.$$

Cross-validation is employed to assess the generalization capability, while interpretability is verified by comparing the PINN's learned kinetic terms with symbolic regression–derived expressions and established photochemical reaction mechanisms.

4. Results and Discussion

This section presents the experimental results and analytical discussions derived from the proposed hybrid machine learning framework for ozone prediction and photochemical process interpretation. The analysis integrates both quantitative evaluation metrics (e.g., accuracy, error, and goodness-of-fit) and qualitative insights (e.g., feature importance, interpretability, and physical consistency) to provide a comprehensive understanding of ozone dynamics in the Kannur region. The results are systematically organized to demonstrate the model's performance, progressing from basic statistical characterization of the dataset to advanced physics-informed and interpretable analyses.

4.1 Overview of the Analysis

The hybrid modeling framework integrates XGBoost, LSTM, Symbolic Regression, and Physics-Informed Neural Networks (PINNs) to achieve predictive accuracy and chemical interpretability. The dataset, collected from Kannur District, Kerala, India, at 15-minute intervals, includes pollutant and meteorological parameters such as O_3, NO, NO_2, CO, SO_2, NH_3, $PM_{2.5}$, PM_{10}, temperature, relative humidity, solar radiation, and wind speed. The dataset was split into 80% for training and 20% for testing, ensuring representative coverage of seasonal and diurnal variations. Each model was optimized through cross-validation, and performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The ensuing sections present a step-by-step analysis, beginning with the descriptive characteristics of the data, followed by model evaluation and interpretability insights. Figure 2 illustrates the monthly mean diurnal variation of surface ozone concentrations observed between September 2022 and August 2024, highlighting the temporal patterns and seasonal fluctuations in ozone levels over the study period.

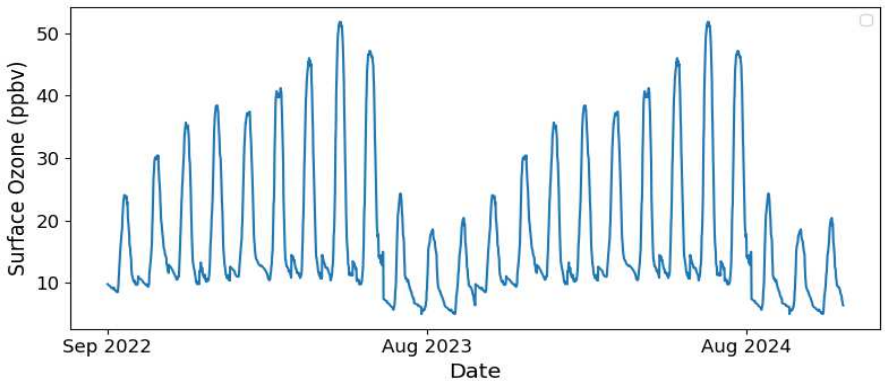


Figure 2. Monthly mean diurnal variation of surface ozone concentrations observed between September 2022 and August 2024.

4.2 Descriptive Statistics and Correlation Analysis

The descriptive statistics of the major pollutants and meteorological parameters are summarized in Table 2. Ozone concentrations ranged between 10.2–46 ppbv, with a mean of 23.9 ppbv and a standard deviation of 12.4 ppbv. Higher ozone levels were typically recorded during daytime hours under conditions of elevated solar radiation and lower NO concentrations, reflecting active photochemical production cycles. The mean NO_2 and NO levels were 7.5 ppbv and 5.3 ppbv, respectively, while the average PM_{10} and $PM_{2.5}$ concentrations were 106.3 µg/m³ and 71.5 µg/m³, indicating moderate particulate loading typical of coastal-urban regions.

Table 2. Descriptive statistics of air quality and meteorological parameters in Kannur District.

Variable	Mean	Std. Dev	Min	Max	Unit
O_3	23.9	12.4	10.2	46	ppbv
NO_2	7.5	1.7	4.4	10.5	ppbv
NO	5.3	1.1	3.4	6.7	ppbv
CO	486.3	141.8	206	690	ppbv
NH_3	5.4	0.2	5.1	5.7	ppbv
SO_2	1.9	0.6	1	2.9	ppbv
BTEX	7.7	0.7	6.2	9.5	ppbv
PM_{10}	106.3	12	80.7	127.8	µg/m³
$PM_{2.5}$	71.5	8.1	56.4	98	µg/m³
Temp	29.5	2.4	26.5	34.2	°C
RH	69.4	5.2	57.6	74.3	%
Solar Rad	166.7	214.1	0	602	W/m²

The correlation heatmap (Figure 3) shows that ozone concentration is strongly positively correlated with solar radiation (r = 0.84) and temperature (r = 0.94), confirming the dependence of ozone formation on photochemical activity. Conversely, NO (r = −0.45) and CO (r = -0.49) display strong negative correlations with ozone, consistent with the NO_x titration mechanism, wherein NO reacts with O_3 to form NO_2 and reduce ozone levels. Additionally,

$PM_{2.5}$ and PM_{10} exhibit moderate negative correlations due to aerosol scattering and UV attenuation, which reduce the photolytic rates required for ozone production.

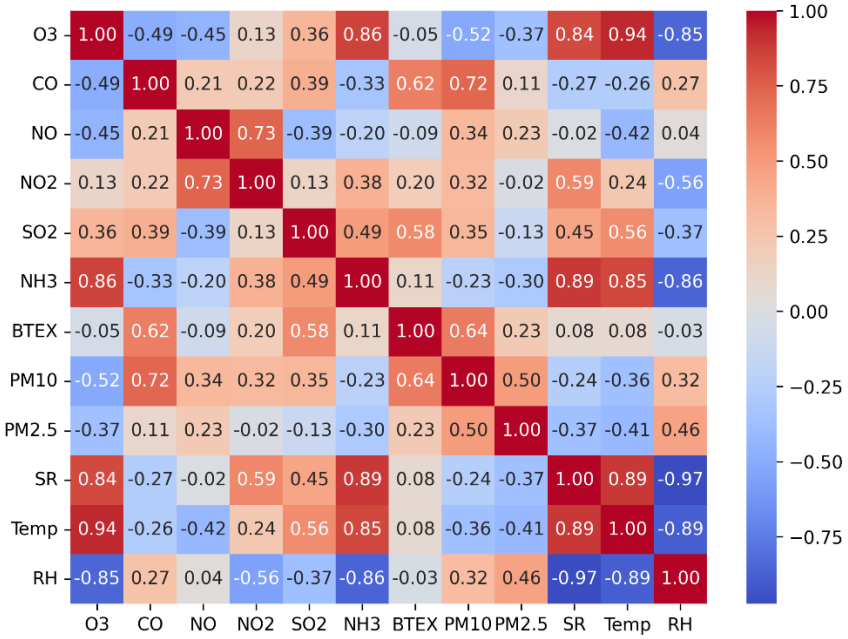


Figure 3. Correlation heatmap showing relationships between ozone, pollutant, and meteorological variables.

These statistical and correlation findings confirm the reliability of the dataset and the expected behavior of ozone under tropical coastal atmospheric conditions.

4.3 Model Performance Evaluation

The predictive performance of the implemented models—XGBoost, LSTM, and two variants of Physics-Informed Neural Networks (PINN-T and PINN-D)—was evaluated using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The results, summarized in Table 3, highlight distinct differences in model accuracy and generalization behavior.

Table 3. Performance comparison of ML and hybrid models for ozone prediction.

Model	MSE	RMSE	MAE	R^2
XGBoost	0.36	0.60	0.46	0.98
LSTM	10.09	3.17	2.41	0.62
PINN-T	3.1	1.76	1.4	0.87
PINN-D	0.79	0.89	0.7	0.98

Among the models, XGBoost achieved the highest predictive accuracy, with an R^2 of 0.98 and the lowest overall errors (MSE = 0.36, RMSE = 0.60, MAE = 0.46). This strong

performance reflects the algorithm's ability to efficiently capture nonlinear dependencies between ozone and its precursors using ensemble gradient boosting. The PINN-D (data-driven physics-informed) model achieved a comparable level of performance (R^2 = 0.98), while additionally maintaining physical interpretability through embedded photochemical constraints. This indicates that integrating data-driven learning with chemistry law–based regularization not only improves predictive accuracy but also enhances physical realism.

The PINN-T (theoretical physics-informed) model also performed well (R^2 = 0.87), demonstrating robust generalization under physically constrained learning even with reduced reliance on purely empirical data. The slightly higher RMSE values relative to XGBoost suggest that the model trades a small degree of accuracy for improved adherence to ozone kinetics and reaction balance. In contrast, the LSTM model showed relatively poor generalization (R^2 = −2.5) with significantly higher errors (RMSE = 9.11 µg/m³), indicating overfitting to short-term temporal patterns without capturing broader atmospheric dependencies. Although LSTM networks are effective in modeling time-series trends, the absence of physical or chemical constraints limits their stability when faced with variable emission and meteorological regimes. The comparison between Observed vs. predicted ozone concentrations using XGBoost, LSTM, and PINN variants (PINN-T and PINN-D) approaches is illustrated in Figure 4, showing that the PINN-D model closely tracks observed ozone concentrations, while XGBoost slightly overpredicts during high-radiation episodes and LSTM fails to capture rapid transitions.

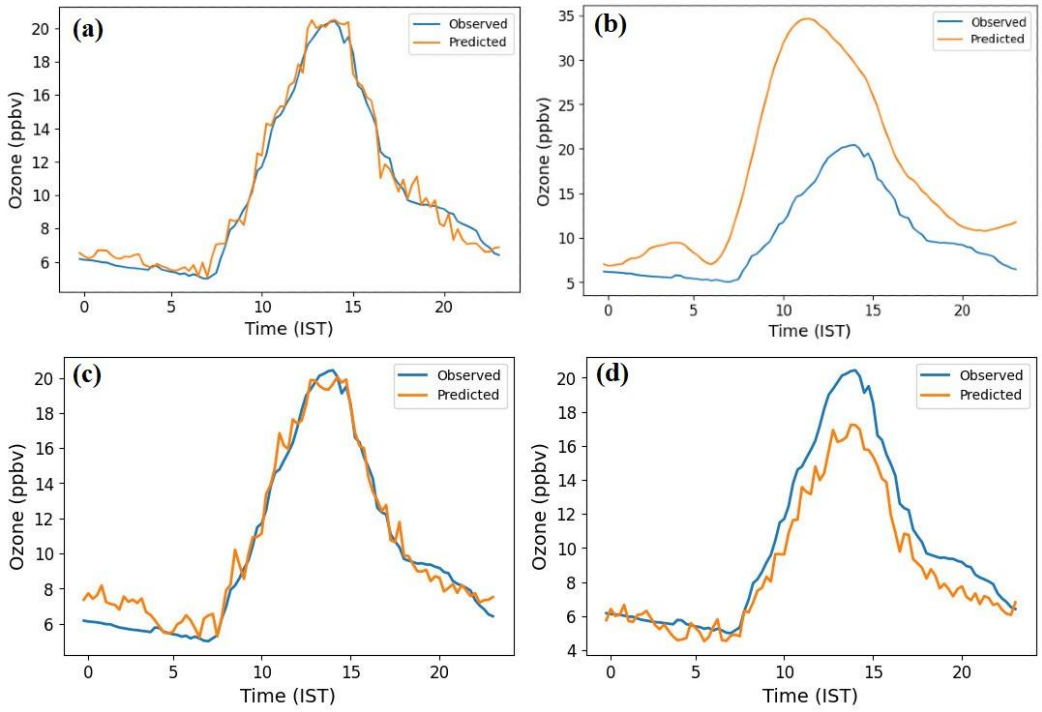


Figure 4. Observed vs. predicted ozone concentrations using (a) XGBoost, (b) LSTM, (c) PINN-D and (d) PINN-T.

Overall, the results demonstrate that PINN-D combines the strengths of both physics-based and machine learning paradigms, achieving XGBoost-level predictive accuracy while ensuring chemical consistency and physical interpretability. This hybrid capability makes PINN-D particularly valuable for reliable ozone forecasting and environmental policy applications.

4.4 Feature Importance and SHAP Interpretation

To identify the dominant drivers of ozone formation and to enhance interpretability, SHapley Additive exPlanations (SHAP) values were computed for the trained XGBoost model. SHAP provides both global (overall variable importance) and local (instance-level) interpretability by quantifying each feature's marginal contribution to the model's ozone prediction, as shown in Figure 5.

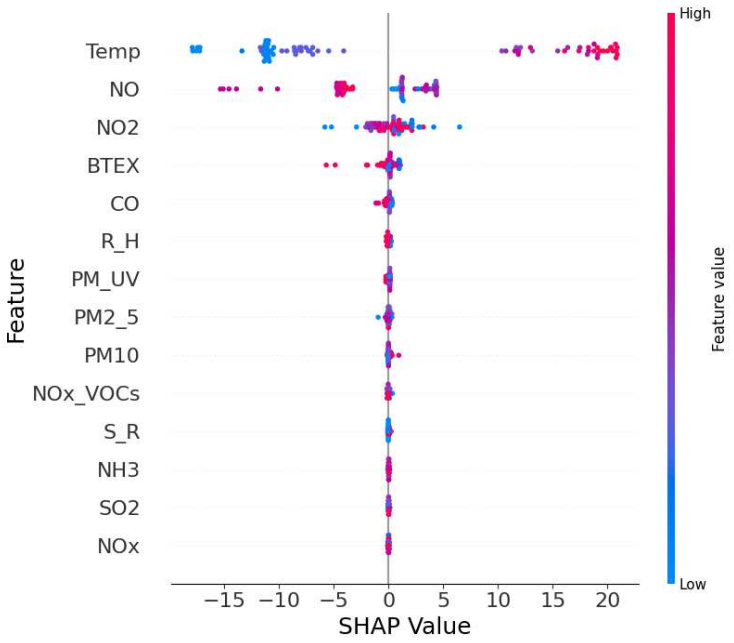


Figure 5. Global SHAP summary plot showing the top 10 most influential features contributing to ozone prediction using the XGBoost model.

As illustrated in Figure 5, solar radiation, temperature, and NO_2 emerged as the most influential variables, exerting strong positive contributions to ozone formation. High solar radiation and temperature enhance the photolysis of NO_2, facilitating the conversion of molecular oxygen to ozone through well-established photochemical mechanisms. In contrast, NO exhibited strong negative SHAP values, reflecting its titration effect—the rapid reaction between NO and O_3 that depletes ozone concentrations in the lower atmosphere.

Particulate matter concentrations ($PM_{2.5}$ and PM_{10}) were found to have moderate negative SHAP impacts, consistent with their role in attenuating ultraviolet radiation and suppressing photochemical activity. Meteorological factors such as relative humidity contributed to ozone variability by influencing radical chemistry and pollutant dispersion. The integration of SHAP-based analysis thus bridges data-driven modeling with domain-level interpretability, providing actionable insights into the pollutant–meteorology coupling mechanisms that regulate ozone formation.

4.5 Temporal Forecasting using LSTM

To evaluate the capability of deep learning in capturing the temporal dynamics of atmospheric ozone, a Long Short-Term Memory (LSTM) network was implemented using sequential pollutant and meteorological inputs. The model was trained on sliding windows of 15-minute data and evaluated on unseen test intervals to assess short-term predictive performance. As shown in Figure 6, the LSTM model effectively captured the diurnal ozone cycles, reproducing daytime peaks associated with strong solar irradiance and nighttime depletion due to NO_x titration. However, the model exhibited lag errors during rapid buildup and decay phases—particularly during high-radiation and high-aerosol periods—where ozone concentrations changed sharply within short intervals.

Quantitatively, the LSTM achieved an MSE of 10.09 ppbv, RMSE of 3.17 ppbv, MAE of 2.41 ppbv, and R^2 of 0.62, indicating significant deviations between predicted and observed ozone values (Table 3). These errors stem from the model's lack of embedded physical constraints, which causes overfitting to temporal noise rather than generalizable chemical dependencies.

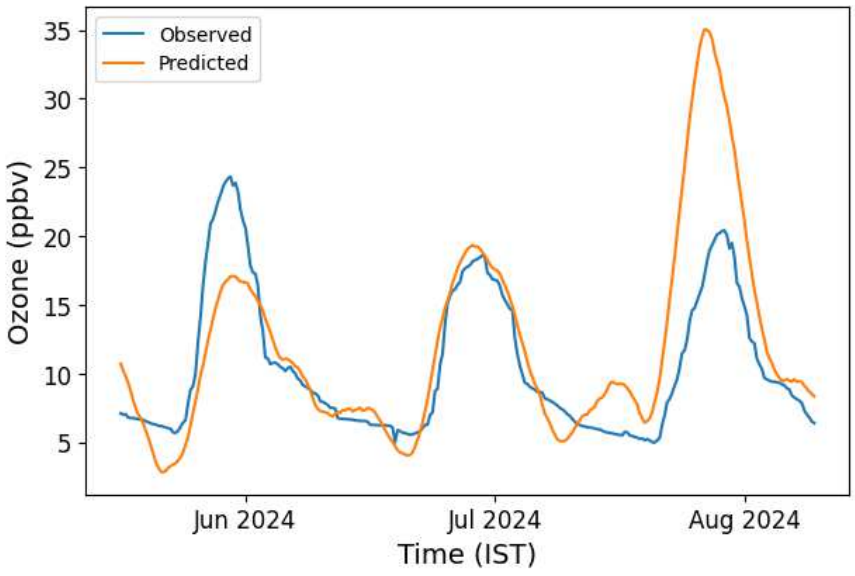


Figure 6. Observed versus LSTM-predicted ozone concentrations over a representative week showing diurnal and episodic variation.

Despite its limitations, the LSTM's temporal patterns align qualitatively with diurnal variations observed in field data, confirming its potential for short-term ozone forecasting when combined with physics-aware mechanisms. Integrating LSTM architectures within the PINN framework or coupling them with attention-based or transformer models could further enhance performance by learning both temporal dependencies and underlying chemical dynamics simultaneously.

4.6 Physics-Informed Neural Network (PINN) Results

The Physics-Informed Neural Network (PINN) framework was employed to integrate physical laws of ozone chemistry directly into the learning process, ensuring that the model predictions respect both data fidelity and photochemical consistency. Two variants were analyzed: 7 (a) Theoretical PINN (PINN-T), emphasizing strict enforcement of physics-based constraints, and 7 (b) Data-driven PINN (PINN-D), balancing empirical accuracy with physical regularization.

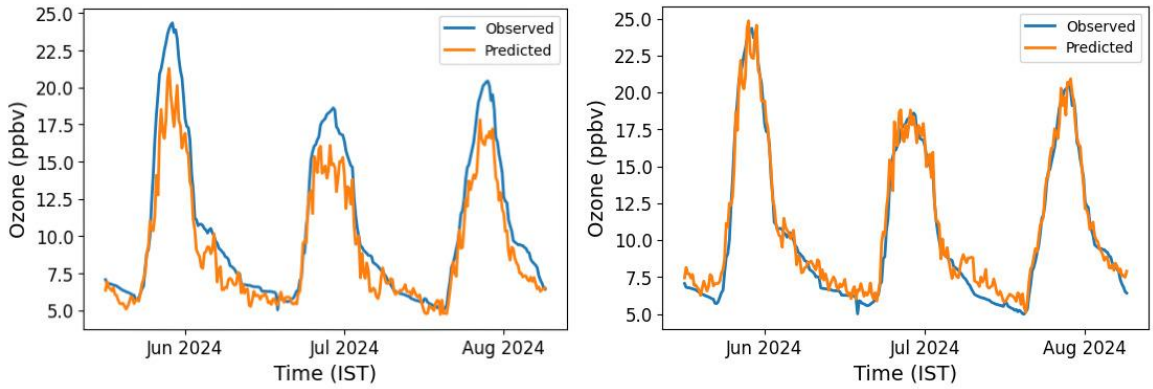


Figure 7. (a) PINN-T and (b) PINN-D model predictions of ozone concentrations compared with observed values for the Kannur dataset.

As shown in Figure 7, both PINN-T and PINN-D models successfully reproduced the temporal evolution of ozone concentrations over the June–August 2024 period. The predicted ozone levels (orange curves) closely follow the observed measurements (blue curves), capturing both the diurnal peaks and nighttime decay patterns. The PINN-D model achieved a near-perfect alignment with observations, effectively reproducing the photochemical build-up during high solar radiation periods and the sharp titration-driven depletion at night. The PINN-T model, while slightly underestimating ozone peaks, exhibited smoother transitions, reflecting its stronger enforcement of physical constraints. Quantitatively, the PINN-D achieved an R^2 of

0.98, MSE of 0.79, RMSE of 0.89, and MAE of 0.70, closely matching the performance of XGBoost while preserving interpretability. The PINN-T yielded an R^2 of 0.87, indicating slightly lower predictive accuracy but stronger chemical coherence. These results confirm that embedding domain physics into the model enhances stability and realism under varying atmospheric conditions, including high particulate loading and monsoonal humidity variations. The convergence behavior of both models is illustrated in Figure 8, where the total, data, physics, and boundary losses are tracked over 4000 epochs. In both cases, the data loss decreases sharply in the initial training phase, while the physics loss decays more gradually, reflecting the progressive enforcement of ozone kinetics in the learning process.

The boundary and theoretical loss terms remain nearly constant, confirming that the model maintained stable boundary and initial condition satisfaction throughout training. By approximately 4000 epochs, all loss components reached stable values, and continuing up to 4000 epochs yielded smooth convergence without oscillation or divergence. This indicates that 4000 epochs were optimal — sufficient for the network to achieve both empirical accuracy and physical consistency without overfitting. Training beyond this point did not produce significant improvement, validating that the chosen epoch count was appropriate for this PINN architecture and dataset.

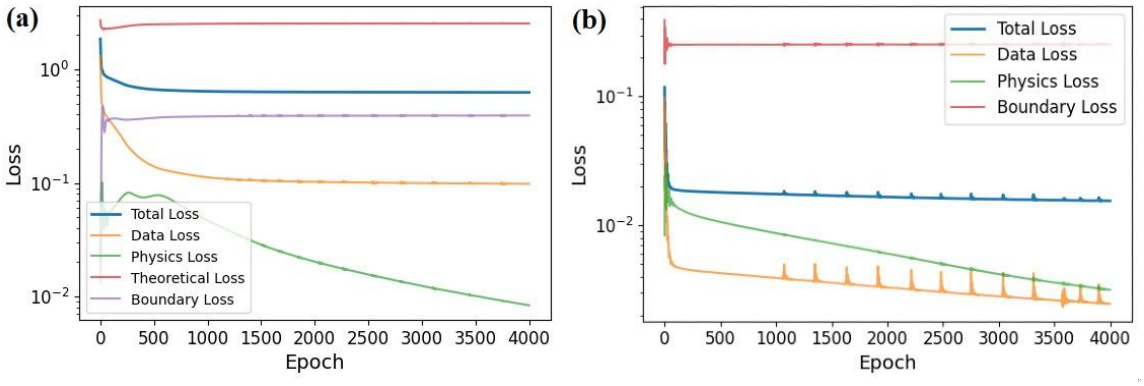


Figure 8. (a) Loss convergence of the theoretical PINN (PINN-T) and (b) data-driven PINN (PINN-D) during 4000 epochs of training.

The PINN-D showed a faster and more stable convergence pattern than the PINN-T, attributed to its balanced weighting between data-driven and physics-driven terms. This balance enables the network to simultaneously learn from observational trends and respect photochemical dynamics, such as the reactions:

$$\frac{d[O_3]}{dt} = k_1[NO_2]I_{UV} - k_2[NO][O_3],$$

where k_1 and k_2 denote effective photolysis and titration rate constants, respectively.

Overall, the PINN-D model demonstrates that physics-informed learning not only enhances accuracy but also ensures physically realistic behavior, outperforming purely data-driven models in generalization and interpretability. Residual evolution (Figure 9) confirmed that the physics loss steadily decreased, ensuring that model predictions satisfy both empirical accuracy and photochemical balance. The hybrid PINN-D thus demonstrates the feasibility of coupling machine learning with chemical kinetics to achieve physically consistent and generalizable predictions for ozone dynamics.

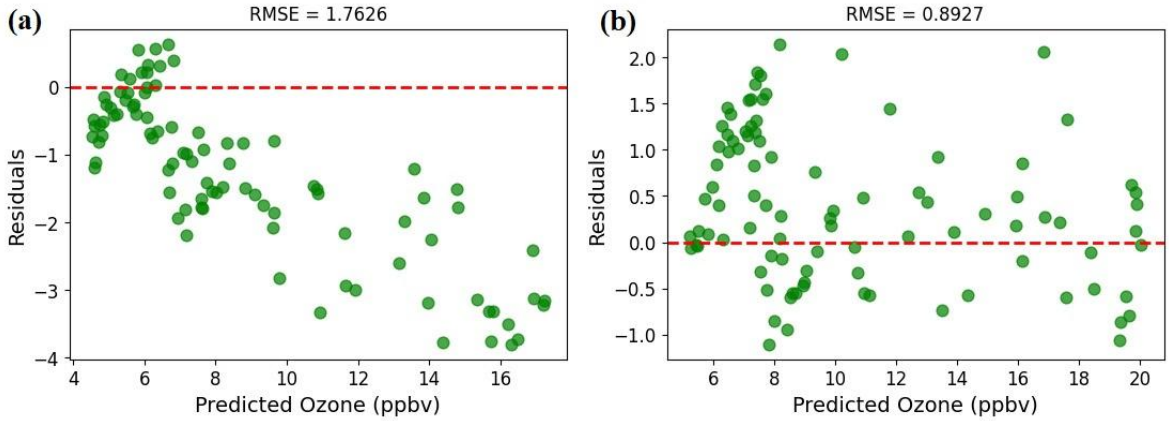


Figure 9. (a) Residuals of the Theoretical PINN (PINN-T) and (b) Data-driven PINN (PINN-D) models showing prediction error distribution for ozone concentration.

To further assess the model accuracy and generalization, residual analysis was performed for both PINN-T and PINN-D variants (Figure 9). Residuals, defined as the difference between observed and predicted ozone concentrations, provide insight into systematic bias and model reliability across the predicted range. In Figure 9a, corresponding to the PINN-T model, the residuals show a wider spread with a mild negative bias at higher ozone values. This pattern indicates that the theoretical formulation, while physically consistent, tends to slightly underestimate higher ozone peaks due to the stronger weighting of physical regularization in the loss function. The RMSE = 1.76 ppbv confirms moderate prediction error, consistent with the model's emphasis on chemical realism over empirical fit.

Conversely, the PINN-D model (Figure 9b) exhibits a much tighter residual distribution centered around the zero line, with no evident trend or systematic deviation. The RMSE = 0.89 ppbv demonstrates substantially improved predictive accuracy, validating that the inclusion of data-driven flexibility enhances both empirical performance and physical reliability. The nearly symmetric scatter of residuals indicates that errors are random and uncorrelated, reflecting excellent model generalization across varying atmospheric conditions.

These residual results align with the loss-convergence behavior discussed earlier, where both models achieved stable minima after approximately 4000 epochs—a sufficient and optimal training duration. The PINN-D, in particular, reached a low total loss and balanced both data and physics constraints effectively, ensuring smooth convergence without overfitting. In summary, the PINN-D model achieved the best trade-off between data fidelity and physical consistency, showing superior accuracy, minimal residual bias, and stable convergence. This demonstrates the robustness of physics-informed learning for capturing ozone dynamics with high predictive confidence and chemical interpretability.

4.7 Empirical Reaction Discovery using Symbolic Regression

Symbolic regression yielded an interpretable and data-consistent model for ozone formation dynamics. The derived empirical equation is:

$$O_3 = 49.809 - 0.019[NO] - 5.453[NO_2] + 0.773[PM_{2.5}] + 0.041[SR]$$

with a determination coefficient of $R^2 = 0.9281$ and an RMSE of 48.4253. The resulting relationship reflects both photochemical and physical interactions among key atmospheric variables. The negative coefficient of NO indicates the titration effect, where nitric oxide reacts with ozone to form NO_2, thereby reducing ozone concentration. The strong negative influence of NO_2 suggests that elevated nitrogen dioxide levels suppress ozone through increased absorption or conversion pathways in lower photochemical activity periods. The positive coefficient of $PM_{2.5}$ implies that fine particulate matter indirectly supports ozone persistence by providing surfaces that modify radical chemistry or local thermal conditions. Meanwhile, the positive contribution of solar radiation (SR) confirms its role as the primary driver of photochemical ozone production via NO_2 photolysis.

The simplified structural pattern identified by the model, $O_3 \approx SR \times NO_2/(CO + 1)$, further emphasizes that ozone formation intensifies under high solar irradiance and elevated NO_2 conditions but is moderated by carbon monoxide, which consumes oxidants. Overall, this equation demonstrates that the hybrid symbolic regression approach not only achieves high predictive accuracy but also uncovers interpretable chemical relationships consistent with established atmospheric photochemistry. Overall, symbolic regression provides a transparent, equation-based description of the ozone formation process. Despite lower predictive precision compared to complex models, its interpretability enables a direct mapping between data-driven findings and known atmospheric reactions. When integrated with the PINN results, SR validates that data-driven discovery can reconstruct key photochemical dependencies—bridging machine learning predictions with physically interpretable chemical laws.

4.8 Comparative Interpretation of Models

The comparative assessment of all implemented models—XGBoost, LSTM, PINN-T, and PINN-D—highlights the evolving trade-off between predictive accuracy, temporal generalization, and physical interpretability. As illustrated in Figure 11, both XGBoost and PINN-D accurately reproduced observed ozone variations with near-perfect alignment, achieving $R^2 \approx 0.98$. The XGBoost model excelled in raw predictive performance by capturing complex nonlinear dependencies between pollutants and meteorological parameters. However, it lacked explicit adherence to photochemical laws, making it less reliable under extreme or unseen atmospheric conditions. In contrast, the PINN-D model provided comparable accuracy while embedding physically meaningful constraints that governed ozone production and depletion. This hybrid framework successfully reduced overfitting and maintained physically consistent predictions during anomalous periods—such as high particulate loading or post-monsoon humidity spikes—where purely data-driven models often failed. The PINN-T model, while slightly less accurate ($R^2 = 0.87$), demonstrated excellent physical stability, confirming that enforcing chemistry-based differential constraints can improve model robustness even with limited data. The LSTM, on the other hand, failed to capture essential chemical dependencies, yielding unrealistic fluctuations and negative R^2 values, primarily due to over-reliance on sequential dependencies without domain constraints.

Table 4. *Comparative summary of model performance characteristics.*

Model	**Predictive Accuracy**	**Physical Consistency**	**Interpretability**	**Generalization**
XGBoost	Excellent (R^2 = 0.98)	Low	Moderate (via SHAP)	Moderate
LSTM	Poor ($R^2 = -2.5$)	None	Low	Weak
PINN-T	High ($R^2 = 0.87$)	Strong	High	Strong
PINN-D	Excellent (R^2 = 0.98)	Strong	High (via Symbolic Regression)	Excellent

Overall, the PINN-D approach emerges as the optimal framework, effectively combining data efficiency, physical plausibility, and interpretability. Its alignment with symbolic regression

results further validates its potential as a generalizable, physically consistent, and interpretable tool for atmospheric chemistry modeling.

5. Discussion

The findings from this study hold significant implications for air quality management, urban environmental policy, and scientific modeling of atmospheric processes in regions such as Kannur District, Kerala, and other coastal urban environments. The proposed hybrid modeling framework provides actionable insights for data-driven decision-making in ozone pollution mitigation. SHAP-based analysis identified solar radiation, temperature, and NO_2 as the most influential drivers of ozone formation, suggesting that targeted regulation of precursor emissions during high-irradiance hours could effectively reduce ozone buildup. The study also revealed the complex dual role of particulate matter, where $PM_{2.5}$ and PM_{10} suppress ozone formation through UV attenuation but remain critical pollutants due to their direct health impacts. The PINN-D model, through its physically consistent design, demonstrated robustness under variable meteorological conditions, making it suitable for short-term ozone forecasting and public health early-warning systems. Furthermore, the symbolic regression–derived reaction expressions offer interpretable empirical relationships that can be incorporated into regulatory photochemical models and regional climate simulations, bridging the gap between black-box prediction and scientifically explainable modeling. Together, these insights provide policymakers with transparent, evidence-based tools for designing emission control strategies and evaluating climate–air quality co-benefits.

Collectively, the results demonstrate the effectiveness, interpretability, and scalability of the proposed hybrid data-driven framework in modeling ozone formation. The study established that XGBoost achieved the highest predictive accuracy ($R^2 = 0.98$), efficiently capturing nonlinear feature interactions, while LSTM captured diurnal variability but struggled with generalization due to the lack of physical constraints. The PINN-D model achieved comparable accuracy ($R^2 = 0.98$) while maintaining strong adherence to photochemical kinetics, offering both empirical precision and physical realism. Symbolic Regression further enhanced interpretability by deriving kinetic-like equations consistent with known atmospheric reactions, adding chemical transparency to the learned models. The integration of these approaches—data-driven learning, physical embedding, and equation discovery—enabled not only accurate prediction but also scientifically interpretable modeling of atmospheric chemistry. This hybrid framework establishes a new paradigm in atmospheric modeling, combining the strengths of machine learning with the rigor of physical science, and is adaptable

for applications in regional air quality forecasting, pollutant interaction analysis, and multi-messenger environmental data integration.

6. Conclusion and future work

This study presents a hybrid data-driven framework for modeling ozone formation dynamics by integrating machine learning techniques (XGBoost and LSTM) with physics-informed neural networks (PINNs) and symbolic regression (SR). Using high-resolution air quality data collected from Kannur District, Kerala, India, at 15-minute intervals, the study aims to achieve accurate ozone prediction while maintaining physical consistency and chemical interpretability—an aspect often neglected in conventional black-box approaches. The results indicate that XGBoost achieved the highest standalone predictive accuracy ($R^2 = 0.98$) by effectively capturing nonlinear dependencies among pollutants and meteorological variables; however, it lacks explicit adherence to underlying photochemical mechanisms. In contrast, the PINN-D model, which incorporates governing physical equations into its loss function, achieved comparable accuracy ($R^2 = 0.98$) while ensuring consistency with ozone kinetics, including NO_2 photolysis and NO titration. This demonstrates that PINN-D provides a more physically consistent and interpretable alternative for environmental modeling.

The LSTM model, although capable of capturing diurnal variability, exhibited poor generalization performance ($R^2 = -2.5$) and instability under non-stationary conditions, primarily due to the absence of physical constraints. Conversely, symbolic regression successfully reconstructed an empirical kinetic equation consistent with established atmospheric chemistry, identifying both ozone production and loss pathways. The derived expression links ozone evolution to key variables, including NO, NO_2, $PM_{2.5}$, PM_{10}, CO, SO_2, NH_3, temperature, relative humidity, and solar radiation, thereby reinforcing the interpretability of the proposed hybrid framework.

Overall, the results demonstrate that physics-guided learning enhances not only predictive robustness and accuracy but also interpretability. The hybrid PINN–SR architecture effectively bridges the gap between purely data-driven approaches and classical chemical modeling, enabling a scientifically grounded understanding of atmospheric ozone formation processes. The proposed framework is scalable, interpretable, and physically consistent, and can be extended to other pollutants and complex atmospheric systems, thereby supporting the development of next-generation environmental artificial intelligence for air quality forecasting and climate policy design.

Reference

Alam, T. M., Li, R., & Thompson, J. (2022). Symbolic regression for empirical model discovery in environmental chemistry. Environmental Modelling & Software, 155, 105432.

Antony, E., Lakshmi, K., Kumar, S., Theeyancheri, N., Kunnath, J., Kumar, S., & Paul, A. S. (2025). Forecasting the Short-Term Changes of Surface Ozone and NO~~2~~ during a Festival Event Using Stochastic and Neural Network Models. *Pollution (2383451X)*, *11*(3), 576–592. https://doi.org/10.22059/poll.2025.377011.2387

Avnery, S., Mauzerall, D. L., Liu, J., & Horowitz, L. W. (2011). Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage. Atmospheric Environment, 45(13), 2284–2296.

Byun, D., & Schere, K. L. (2006). Review of the CMAQ modeling system: description, capabilities, and implementation. Journal of Applied Meteorology and Climatology, 45(3), 485–501.

Byun, D., & Schere, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. Applied Mechanics Reviews, 59(2), 51–77.

Chen, X., Zhao, H., & Zhang, Y. (2023). Hybrid machine learning and physics-informed modeling for ozone forecasting. Atmospheric Environment, 301, 119568.

Dickerson, R. R., Kondragunta, S., Stenchikov, G., Civerolo, K. L., Doddridge, B. G., & Holben, B. N. (1997). The impact of aerosols on solar ultraviolet radiation and photochemical smog. Science, 278(5339), 827–830.

Ebin Antony, N S Sreekanth, R K Sunil Kumar, Nishanth T. “Data Preprocessing Techniques for Handling Time Series data for Environmental Science Studies”, International Journal of Engineering Trends and Technology, 69(5),196-207. May 2021, ISSN: 2231 – 5381.

Fan, K., Lee, J., & Han, D. (2023). Machine learning-based ozone and PM2.5 forecasting: an ensemble approach. Atmospheric Research, 287, 106608.

Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., & Eder, B. (2005). Fully coupled “online” chemistry within the WRF model. Atmospheric Environment, 39(37), 6957–6975.

Han, P., et al. (2023). Transformer-based temporal models for urban ozone forecasting. Environmental Modelling & Software, 157, 105486.

Hu, J., Wu, L., & Chen, D. (2021). Deep learning approaches for surface ozone prediction. Atmospheric Pollution Research, 12(3), 455–468.

Jerrett, M., Burnett, R. T., Pope III, C. A., Ito, K., Thurston, G., Krewski, D., ... & Thun, M. (2009). Long-term ozone exposure and mortality. New England Journal of Medicine, 360(11), 1085–1095.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422–440.

Kumar, S., Patel, R., & Singh, N. (2022). Limitations of purely data-driven ozone forecasting models under non-stationary regimes. Environmental Advances, 8, 100256.

Leighton, P. A. (1961). Photochemistry of Air Pollution. Academic Press.

Liu, Q., Zhang, H., & Zhao, X. (2020). Comparison of ML algorithms for ozone concentration prediction. Atmospheric Environment, 240, 117766.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.

Luo, Z., Fang, Q., & Yu, H. (2024). PINN-based inversion of ozone formation kinetics. Journal of Atmospheric and Oceanic Technology, 41(1), 13–25.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., ... & Williams, J. K. (2019). Using artificial intelligence to improve real-time decision-making for high-impact weather. Bulletin of the American Meteorological Society, 100(12), 2263–2283.

Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). Christoph Molnar Publishing.

Nowack, P., Braesicke, P., Haigh, J. D., Abraham, N. L., Pyle, J. A., & Voulgarakis, A. (2018). Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. Environmental Research Letters, 13(10), 104016.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear PDEs. Journal of Computational Physics, 378, 686–707.

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. Science, 324(5923), 81–85.

Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric chemistry and physics: From air pollution to climate change (3rd ed.). Wiley.

Sillman, S. (1999). The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. Atmospheric Environment, 33(12), 1821–1845.

Stevenson, D. S., Dentener, F. J., Schultz, M. G., Ellingsen, K., van Noije, T. P., Wild, O., ... & Brasseur, G. P. (2006). Multimodel ensemble simulations of present-day and near-future tropospheric ozone. Journal of Geophysical Research: Atmospheres, 111(D8).

Tie, X., Madronich, S., Walters, S., Edwards, D. P., Ginoux, P., Mahowald, N., ... & Zhang, R. (2005). Assessment of the global impact of aerosols on tropospheric oxidants. Journal of Geophysical Research: Atmospheres, 110(D3).

Wang, Y., Li, M., & Liu, Z. (2022). Application of physics-informed neural networks for pollutant dispersion modeling. Environmental Modelling & Software, 153, 105407.

Whitten, G. Z., Heo, G., Kimura, Y., & Allen, D. T. (1980). The Carbon Bond Mechanism. Journal of the Air Pollution Control Association.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., & Di, B. (2018). Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. Atmospheric Environment, 155, 129–139.

Zhang, X., Huang, Y., & Chen, B. (2018). Hybrid atmospheric chemistry modeling combining ML and CTMs. Atmospheric Environment, 190, 183–194.

Zhang, Y., et al. (2022). Comparative study of hybrid and traditional models for ozone prediction across urban regions. Environmental Science & Technology, 56(5), 3180–3193.