



RESEARCH ARTICLE

10.1029/2024JD041593

Key Points:

- We introduce NetGBM, a machine-learning model for high spatiotemporal resolution ozone prediction across China
- We perform ozone inhalation and GEMM analysis, which identifies regions facing elevated health risks due to high ozone exposure
- The NetGBM model has superior performance compared to prior research and performs robustly in regions with limited monitoring resources

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Song,
junsong@hkbu.edu.hk

Citation:

Ma, C., Song, J., Ran, M., Wan, Z., Guo, Y., & Gao, M. (2024). Machine learning-driven spatiotemporal analysis of ozone exposure and health risks in China. *Journal of Geophysical Research: Atmospheres*, 129, e2024JD041593. <https://doi.org/10.1029/2024JD041593>

Received 18 MAY 2024

Accepted 7 SEP 2024

Author Contributions:

Conceptualization: Chendong Ma, Jun Song

Data curation: Chendong Ma, Maohao Ran, Zhenglin Wan

Formal analysis: Chendong Ma

Funding acquisition: Jun Song

Investigation: Chendong Ma, Jun Song

Methodology: Chendong Ma, Jun Song

Project administration: Jun Song

Resources: Jun Song

Software: Maohao Ran, Zhenglin Wan

Supervision: Jun Song, Yike Guo, Meng Gao



Validation: Jun Song, Maohao Ran, Zhenglin Wan

© 2024 The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License,

which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Machine Learning-Driven Spatiotemporal Analysis of Ozone Exposure and Health Risks in China

Chendong Ma¹, Jun Song¹ , Maohao Ran¹, Zhenglin Wan², Yike Guo³ , and Meng Gao¹

¹Department of Geography, Hong Kong Baptist University, Hong Kong, China, ²School of Data Science, The Chinese University of Hong Kong, Shenzhen, China, ³Hong Kong University of Science and Technology, Hong Kong, China

Abstract Accurate and fine-scaled prediction of ozone concentrations across space and time, as well as the assessment of associated human risks, is crucial for protecting public health and promoting environmental conservation. This paper introduces NetGBM, an innovative machine-learning model designed to comprehensively model ozone levels across China's diverse topography and analyze the spatiotemporal distribution of ozone and exposure. Our model focuses on daily, weekly, and monthly predictions, achieving commendable R^2 coefficients of 0.83, 0.77, and 0.79, respectively. By constructing a gridded map of ozone and incorporating both land use and meteorological features into each grid, we achieved ozone prediction at a high spatiotemporal resolution, outperforming previous research in terms of performance and scale, particularly in regions with limited monitoring stations. The results can be further improved when applied to regional research using meteorological and ozone data from regional stations. Additionally, our research revealed that temperature is the most significant factor affecting ozone concentrations across China. In health risk assessment, we retrieved a high-resolution spatial distribution of ozone-attributed mortality for 5-COD and daily ozone inhalation distributions during our study period. We concluded that ozone-attributed mortality is predominantly caused by stroke and IHD, accounting for more than 70% of the total deaths in 2021, with the highest mortality rates in developed urban areas such as the NCP and the YRD. Our experiment demonstrated the potential of NetGBM in robustly modeling ozone across China with high spatiotemporal resolution and its applicability in measuring associated health risks.

Plain Language Summary This study introduces NetGBM, an innovative machine-learning model designed to forecast ozone levels and assess ozone-attributed public health risks. This model is crucial for post-pandemic air quality management, providing high-resolution predictions that are essential for targeted health interventions and informed environmental policies. By integrating feature engineering with predictive analytics, NetGBM enhances its performance, particularly in regions with limited monitoring. This makes it a robust tool for developing sustainable environmental strategies.

1. Introduction

In the contemporary era, addressing the intricate challenges posed by air pollution has emerged as a paramount concern, particularly in rapidly developing nations such as China. In China, the developments, as well as the COVID-19 pandemic and the associated management strategies, added complex layers to the multifaceted conversation surrounding the air pollution problem (Zheng et al., 2021). Among these pollutants, a crucial but often understudied component of this crisis is elevated levels of tropospheric ozone (O_3), a potent greenhouse gas and a significant contributor to smog and ill health. During the past decade, there has been a notable decrease in most air pollutants in China, attributed to various factors including governmental controls and the influence of the pandemic. However, paradoxically, the levels of ozone pollution have increased substantially. Specifically, there was a 10.79% increase registered from 2013 to 2016, and ozone levels uniquely experienced an unanticipated surge even during the ongoing pandemic period (Chen et al., 2023; Le et al., 2020; G. Yang et al., 2020; S. Zhu et al., 2021). Given China's economic clout and accelerating urbanization, understanding the dynamics of O_3 pollution in this region provides valuable insights into this cutting-edge research domain. This work intends to apply some statistical modeling techniques to model a high-resolution spatial-temporal O_3 concentration across the whole country, as well as the exposure of the Chinese population toward O_3 , based on the meteorological data collected from fixed stations across China, thereby contributing to the growing body of knowledge in this vital area of inquiry.

Visualization: Chendong Ma,
Maohao Ran
Writing – original draft: Chendong Ma
Writing – review & editing:
Chendong Ma, Jun Song

In the realm of previous research, significant strides have been taken toward advancing the modeling of ozone concentrations, primarily focusing on simulation techniques, monitoring networks, and the exploration of pollutant interactions within atmospheric chemistry (J. Li, Nagashima, et al., 2019; Ren et al., 2020). Notably, modeling frameworks such as the Community Multi-scale Air Quality (CMAQ) system and the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) have been emblematic of this approach (J. Hu et al., 2016; Luecken et al., 2019; Tie et al., 2013; Wang et al., 2022). However, it is imperative to acknowledge that most of the research was based on data before 2019 within specific regions or major cities, and that the efficacy of these prominent methodologies heavily relies on the data gleaned from the limited fixed monitoring stations in urban areas. Data sufficiency and lack of granularity remain significant challenges for high spatial-temporal resolution ozone analysis in China (Hong et al., 2019; Requía et al., 2020). Notably, most of the monitoring infrastructure for O₃ in China was only introduced in the more developed areas such as the southeast coast, coupled with a deficiency of monitoring stations in remote and less-studied regions such as the western highlands. On the other hand, although there are traditional global chemistry-climate models (CCMs) that can measure ozone concentrations on a large scale, such as GEOS-Chem, these models measure ozone concentrations across all vertical levels, including both tropospheric and stratospheric ozone, and lack sufficient analysis of ozone concentrations at or near the surface, which is more critical for public health assessment (Henze et al., 2007; Lu et al., 2020). To the best of the authors' knowledge, there is still a lack of methods that analyze ozone trends across China, focusing on surface-level ozone concentrations with high spatial-temporal resolution in recent years. Hence, evaluating and analyzing the long-term effects of air pollution with a high degree of spatial-temporal resolution using the spatially restricted pool of available observations has become a focal point of the present study.

Meteorological and spatial features are vital components in the modeling of ozone concentrations—substances that impact air quality significantly. Atmospheric conditions, like temperature, solar radiation, wind speed, and humidity, can influence the ozone production rate (Abdi-Oskouei et al., 2020; Iglesias-Gonzalez et al., 2020; R. Ma et al., 2021). For instance, high temperatures and solar radiation can accelerate the photo-chemical reactions involving O₃, while wind dispersion can distribute the pollutants on a broader scale (Abdullah et al., 2019; Fu & Tian, 2019). In our previous research, we successfully integrated static and dynamic features in air pollution modeling for Chengdu, China, to enable timely predictions of PM_{2.5} concentrations (Song et al., 2020, 2022). Detailed simulations of these features are, therefore, critical for reliable ozone modeling. Beyond meteorological features, the impact of aerosol optical depth (AOD) in modeling Ozone is a topic that has garnered substantial attention from researchers worldwide due to the complex interaction between aerosols and ozone (Q. Liu et al., 2019). Aerosols are known to reduce surface ozone by lowering the amount of solar radiation reaching the surface, a phenomenon referred to as the “radiative effect.” The magnitude of this impact depends mainly on the distribution, characteristics, and absorbing or scattering nature of the aerosols (T. Zhang et al., 2021). Conversely, aerosols can also enhance ozone levels via the “heterogeneous chemical effect,” acting as a surface for chemical reactions, particularly in the presence of nitrogen oxides and volatile organic compounds (T. Zhang et al., 2021). Thus, AOD plays a significant role in accurately modeling and estimating O₃ levels, making it crucial to incorporate AOD data sets into atmospheric models for improved predictive capabilities (Stafoggia et al., 2020).

In recent years, statistical methodologies have been extensively employed to model O₃ concentrations, yielding noteworthy insights into the intricate analysis of meteorological situations, precursor emissions, and ozone intensity. Multiple Linear Regression (MLR), one of the most recurrently used linear statistical models, is frequently utilized for O₃ estimating, primarily examining linear relationships between environmental ozone concentrations and meteorological attributes alongside precursor gases that contribute to the formation of O₃. Such models have witnessed prediction accuracies varying between 0.3 and 0.75, contingent on selected features and geographical origin (Allu et al., 2020). Non-linear techniques, particularly Neural Networks (NNs), have also been exhibited, yielding a high level of precision ranging between 0.6 and 0.9 due to their proficiency in noise reduction in complicated and non-linear atmospheric chemistries (AlOmar et al., 2020). Some non-parametric approaches, such as kNN and deep neural networks, have been successful in spatial analysis. However, these typically falter in spatio-temporal analysis (Ren et al., 2020). XGBoost has been discovered to have the pinnacle of spatiotemporal performance in Ozone modeling, with a potential of attaining an R² as high as 0.96 by tracing VOCs precursor emissions (Cheng et al., 2023). While these models serve as substantial assets for Ozone modeling and policy development, their effectiveness heavily relies on numerous variables, such as VOCs and the geographic positioning of monitoring equipment. However, VOCs emissions documentation is not always

thoroughly executed due to the limited observation monitors, notably in underdeveloped rural regions, and the distribution of ground Ozone monitors across the nation remains sporadic. Consequently, taking into account these elements, our aspiration is to develop a robust model that offers high resolution in both spatial and temporal spectra and is proficient at functioning effectively on both national and regional-specific scales.

This paper makes two unique contributions to the field of ozone pollution research. Firstly, it presents a comprehensive examination of high spatial-temporal resolution modeling of ozone concentrations using sparsely distributed data. The study addresses the challenges posed by China's diverse topography and varied emission sources, introducing the NetGBM algorithm to enhance the predictive accuracy of ozone concentrations across both spatial and temporal scales. Secondly, the paper applies the ozone modeling results to estimate health impact based on the Global Exposure Mortality Model (GEMM) and ozone exposure among populations. This novel approach bridges the gap between atmospheric modeling and public health, enabling a more accurate assessment of the potential health impacts of ozone pollution.

By combining advanced modeling techniques with population exposure assessment, this study offers valuable insights to guide sustainable air quality management strategies and inform well-informed policy formulations. The findings contribute to a better understanding of the complex dynamics of ozone pollution in China and its potential health consequences. The paper aims to provide a robust framework that can be adapted to other regions facing similar challenges, ultimately promoting evidence-based decision-making in environmental health.

2. Method

2.1. Data Sets

2.1.1. National Model

Our study utilized daily O₃ concentration data collected from 1,733 air quality monitoring stations across China. To facilitate analysis, we divided the national map into 0.25° × 0.25° grids. Within each grid, we calculated the mean ozone concentration by averaging the readings from the stations located within that particular grid. Similarly, we obtained meteorological attributes from weather stations within each grid, computing daily averages for each grid. Additionally, we documented static urban aspects such as green coverage and land use, segmenting them into discrete grids. Each data unit comprises three groups of features: static and dynamic features, location information and time epoch, as depicted in Figures 1b and 1c. For further details on data pre-processing and feature information, please refer to the last section.

2.1.2. Regional Model

In our experiment, to investigate potential performance differences when applying the model to national versus regional data sets, we implemented the model separately for two regions, NCP and YRD. Unlike the national model, we utilized meteorological and ozone data from regional stations within the specific areas. We selected training and test sets from this new data set to better account for the spatial attributes influencing ozone concentrations. The distributions of the labeled grids and the selection of training and test sets are specifically illustrated in Figure S1 in Supporting Information S1. Further details on the regional model, as well as a comparison of results and discussions, can be found in Section 3.2.1.

2.2. NetGBM Architecture

The comprehensive experimental procedure of NetGBM is delineated in Figure 2. The procedure commenced with the partitioning of land use and meteorological data into grid segments, each covering an area of 25 × 25 km, yielding static (V) and dynamic features (V_t), respectively. After this segmentation, the model was applied to these distinct grid segments, each encompassing both static and dynamic features. A rigorous data cleansing protocol preceded this phase, followed by employing a feature engineering approach facilitated by a neural network model. Following the feature engineering process, observations were manipulated by utilizing LightGBM, an advanced technique adeptly tailored for modeling the intricate spatial-temporal distribution of atmospheric pollutants. Combining the neural network process and LightGBM forms our novel NetGBM mechanism. To further extend the analysis, population data was incorporated into the ozone distribution analysis to conduct modeling on ozone exposure. This integration allows for a comprehensive assessment of the potential health impacts associated with ozone pollution.

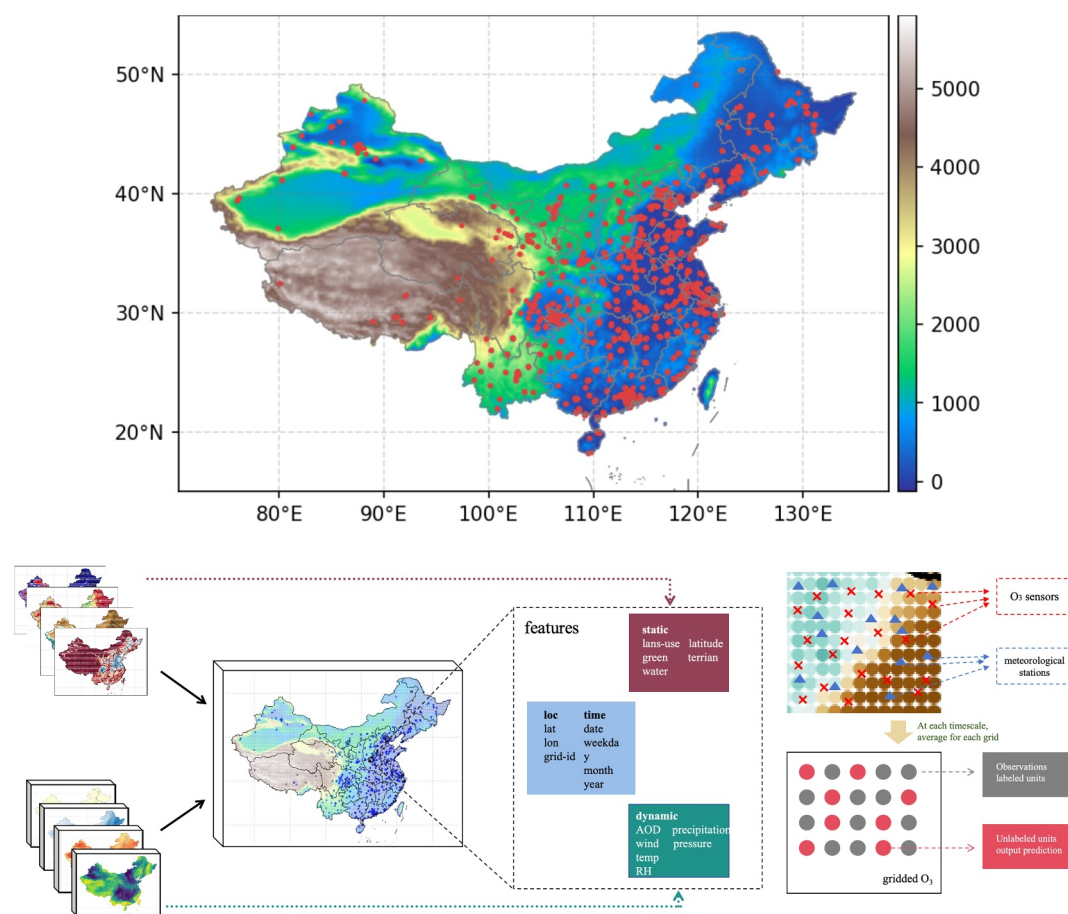


Figure 1. The observation data is structured like this. (a) The distribution of 1,733 national O₃ monitoring stations across China, with color-coded elevation indications overlaid on the map; (b) a scheme of the extracted features for each observation unit; (c) the O₃ concentration and meteorological data for each unit are calculated by averaging the data from corresponding sensors and stations in related grids and time-scales. Any grids without observations are considered unlabeled data and will be the focus of our predictions.

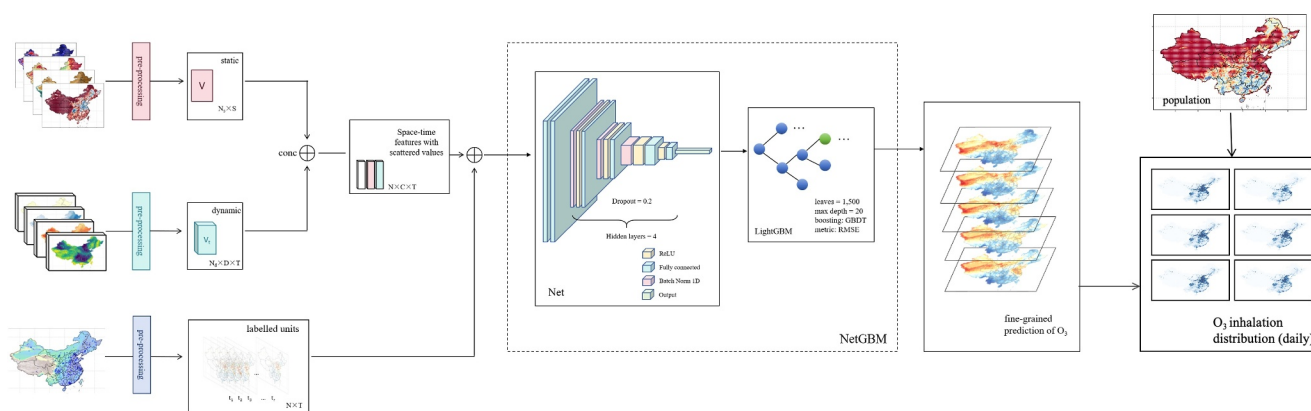


Figure 2. The overall architecture of our model. It shows the incorporation of static and dynamic input values, which are concatenated into a unified feature map. This feature map subsequently undergoes feature engineering, followed by NetGBM training. The resultant model is then employed to predict O₃ concentrations and combine them with the population distribution for the exposure analysis.

To mitigate potential overfitting, a thorough 10-fold cross-validation procedure was meticulously conducted. Upon completion of model training, it was subsequently applied to designated test data sets for predictive analysis. The predictions derived from this process were subsequently juxtaposed against actual observations to ascertain the model's efficacy in predictive accuracy. To quantitatively assess the model's efficacy, an R-squared (R^2) metric was employed as a measure of explanatory power. In summary, our experimental workflow encapsulates a methodical progression from data partitioning to model evaluation, encompassing sophisticated techniques and rigorous evaluation measures to ensure the robustness and precision of our analysis.

2.3. Feature Engineering

In the feature engineering process, we learn which features impact O_3 concentrations by observing labeled data and providing those features to the learning process. There are two types of features we consider: static features, like geographic and land-use information, as well as the spatial-temporal information, denoted by $V \in \mathbb{R}^{N_s \times S}$, where N_s is the number of labeled grids with valid static feature readings, in our case is the number of training grids 697, and $S = 13$ is the number of static features extracted. Secondly, we include meteorological data as dynamic features. Dynamic features differ from static features in that they change over time, so we denote them as $V^t \in \mathbb{R}^{N_d \times D \times T}$, where N_d is the number of labeled grids with valid dynamic feature readings, $D = 8$ is the number of dynamic features, and T is the number of time points. The details about our features can be found in Table 3. In general, this process can be represented as

$$W = V \oplus V^t \quad (1)$$

$X_{static} \xrightarrow{\quad \uparrow \quad} \quad \uparrow \quad \xleftarrow{\quad} X_{dynamic}$

where $W \in \mathbb{R}^{N \times C}$ represents the input volume, N represents the number of observation units and C represents the number of features. Each unit is defined by its spatial location (the specific grid) and temporal point (the date). The information on each unit's location and timestamp are also included in the data set as features. To create the input values, we take the averages of the static features for each grid within the labeled data, resulting in a concatenated static feature input value $V \in \mathbb{R}^{N_s \times S}$. This process can be expressed as:

$$V = \oplus (X_s^{(i)}, i \in \{1, \dots, S\}) \quad (2)$$

where $X_s^{(i)}$ represents each static feature i .

Similarly, for each time point t , we take the averages of the dynamic features within each grid, resulting in a set of dynamic feature input values $\{V^t : t \in T\}$, where $V^t \in \mathbb{R}^{N_d \times D}$ for each t . This process can be represented as:

$$V^t = \oplus (X_d^{(j,t)}, j \in \{1, \dots, D\}) \quad (3)$$

where $X_d^{(j,t)}$ represents the dynamic features j at time point t .

Then, we concatenate V and V^t into one input data set W , which we provide to the neural network process for further learning. The concatenation process can be expressed as:

$$W = \oplus (V, V^t) \quad (4)$$

where \oplus denotes the concatenation operation, resulting in $W \in \mathbb{R}^{(N_s + N_d \times T) \times (S + D)}$.

2.4. Neural Network and LightGBM

The spatial-temporal modeling of O_3 initially employed LightGBM as a primary choice due to its exemplary track record in air pollution modeling tasks. LightGBM's unique feature engineering process and gradient-based sampling methodology have been recognized for their ability to substantially enhance computational efficiency without sacrificing predictive accuracy (Ke et al., 2017). Its successful application in many high-resolution air pollution prediction research studies further advocates its efficacy (Wei et al., 2021; Y. Zhang et al., 2019,

2020; Zhong et al., 2021). Nevertheless, applying LightGBM to our data set did not yield results on par with previous works. This discrepancy could be traced back to two fundamental issues. The first issue lies within the histogram-based approach employed by LightGBM, which can pose challenges in handling sparse data containing missing values. This characteristic is particularly relevant to our data set, which exhibits sparsity across spatial and temporal scales. The second concern arises from the inherent mismatch between LightGBM, a tree-based gradient model, and image processing tasks. Our data set, on the other hand, exhibits a grid-like structure that closely resembles image data.

To counter these issues, rather than resorting to alternative methods that come with their own sets of limitations, we chose to integrate a neural network (Net) feature extraction process before feeding the data into LightGBM for learning. The Net model can be represented as:

$$h = f(Wx + b) \quad (5)$$

where h is the output of the neural network, f is the activation function, W is the weight matrix, x is the input data, and b is the bias term. A Net model contributes sparse regularization, fostering sparsity in the learned weights. This network automatically prioritizes the most relevant features by penalizing large weights and encouraging many weights to approach zero, thereby improving generalization. The sparse regularization can be achieved by adding a regularization term to the loss function:

$$L = L_0 + \lambda \sum_{i,j} |W_{ij}| \quad (6)$$

where L_0 is the original loss function, λ is the regularization coefficient, and $|W_{ij}|$ represents the absolute value of the weight at position (i, j) in the weight matrix. Furthermore, Net models are particularly well-suited for processing image data. After the Net feature extraction, the processed data is more amenable to learning using LightGBM. The LightGBM model can be expressed as:

$$\hat{y} = \sum_{i=1}^T f_i(x) \quad (7)$$

where \hat{y} is the predicted output, T is the number of decision trees, and $f_i(x)$ is the prediction of the i -th tree. Our experimental results showcase an enhanced performance compared to the standalone use of LightGBM while preserving an acceptable computational expense.

2.5. Ozone Exposure and Mortality Assessment

Exposure inhalation mainly causes adverse health effects, and it's important to estimate the inhalation volume in total, which constitutes one of the most crucial preconditions to model the correlations between air pollutant exposure and public health. To estimate the cumulative inhalation volume at a specific area for a certain air pollutant, say task $p \in P$, the inhalation rate for different groups of people, population density and the space-time variations of the air pollutant are fused to achieve a better estimate. The estimation formula of inhalation is shown in the following that, for any task p ,

$$Inh_p = \sum_{i=1}^N \sum_{t=1}^T (Z_{pop}(i) \cdot h_i \cdot d_i \cdot X_p(i, t) + Z_{pop}(i) \cdot h_i \cdot (1 - d_i) \cdot X_p(i, t) \cdot \alpha_p), \forall p \in P \quad (8)$$

where Inh_p denotes the cumulatively inhaled mass of task p (unit: μg) at this specific area, in our case ozone, and h_i denotes the inhalation rate (unit: m^3/h) for the i -th group of the population, which is separated by any valid criteria. N represents the overall amount of those groups, and t denotes the time (days in this study). $X_p(i, t)$ denotes the air pollutant concentration (unit: $\mu\text{g}/\text{m}^3$) for task p at time t with group i , while $Z_{pop}(i)$ denotes the corresponding population (unit: case) in a specific group, T is the target temporal period, d_i denotes the percentage of the population outdoors, and α_p denotes the outdoor-indoor ratio of the concentration for task p .

Specific Equation 8 limitations exist due to their dependency on multiple variables. For example, the ratio of outdoor to indoor exposure to air pollutants is impacted by various details, including geographic location, building structures, etc. Besides, the inhalation rate varies across age, sex, and other factors affecting the inhaled value (Marty et al., 2002). Hence, based on the equation, we simplify the calculation by neglecting other factors and choosing to group the population by spatial locations with our gridded map. Thus, the estimations of cumulative inhaled masses can be directly obtained as follows:

$$Inh'_p = \sum_{i=1}^N \sum_{t=1}^T Z_{pop}(i) \cdot h_p(i) \cdot X_p(i, t), \forall p \in P \quad (9)$$

where Inh'_p denotes the cumulatively inhaled mass of task p from the simplified model, Z_{pop} represents population data, where each i stands for a certain grid. h_p reflects the empirical inhaled volume of each task, and $X_p(i, t)$ denotes the concentration of O_3 on grid i in time t . Duan (2015) summarized the daily air inhalation rates based on previous official research. Accordingly, we take the average of the data into each grid, which is the individual inhalation rate at $16.1 \text{ m}^3/\text{day}$ across China (Duan, 2015). Then we calculate the specific inhalation exposure volumes of O_3 in the required regions.

For the estimation of mortality, we apply the Global Exposure Mortality Model (GEMM) to calculate the premature mortality attributable to ozone exposure across China, as described by the following equation:

$$M = \frac{(RR - 1)}{RR} \times Z_{pop} \times D_k \quad (10)$$

where Z_{pop} represents the population data as described above, and D_k is the baseline number of deaths for each cause of death (COD) k within the population. We consider five major causes of death in this analysis: Chronic Obstructive Pulmonary Disease (COPD), Ischemic Heart Disease (IHD), Lung Cancer (LC), Lower Respiratory Infections (LRI), and stroke. The term $k \in K$ denotes each COD, such that D_k corresponds to the baseline mortality for each specific COD k . The term RR represents the relative risk of mortality at a given ozone concentration, calculated using GEMM as established by Burnett et al. (2018). The relative risk equation is given by:

$$RR(x) = \exp\left(\frac{\theta \log(1 + x/\alpha)}{1 + \exp(-(x - \mu)/\nu)}\right) \quad (11)$$

In this equation, x represents the given concentration of ozone, and the parameters θ , α , μ , and ν define the shape of the exposure-response curve. These parameters are determined based on the specific COD and population characteristics, as established by Burnett et al. (2018), with the exact values provided in Supporting Information S1. Utilizing this calculation, we derive the estimates of ozone-attributable mortality, which are presented in the results section (Figure 7).

3. Results and Discussion

3.1. Predictive Performance

The data set under consideration encompasses a total of 747 grid segments, spanning the entire geographical expanse of the country. This coverage extends from December 2019 to November 2021, yielding a substantial collection of 482,727 observations. A visualization of the prediction of O_3 exposure across China can be found in Figure 3a, which shows a comparison of the predicted O_3 concentrations using GEOS-Chem and NetGBM from 20 to 22 August 2021.

The validation phase of the study encompassed a comprehensive 10-fold cross-validation procedure conducted within the training set. Specifically, the overall coefficient of determination (R^2) attained a value of 0.77, accompanied by a root mean square error (RMSE) of $13.11 \mu\text{g m}^{-3}$. Collectively, these findings highlight a significantly heightened level of accuracy in the model's prognostic capabilities concerning daily O_3 concentrations. Intriguingly, an enhanced model performance is discernible as the temporal scale extends. Upon an examination of weekly O_3 concentrations, comprising a substantial sample size of 69,788 observations, the R^2

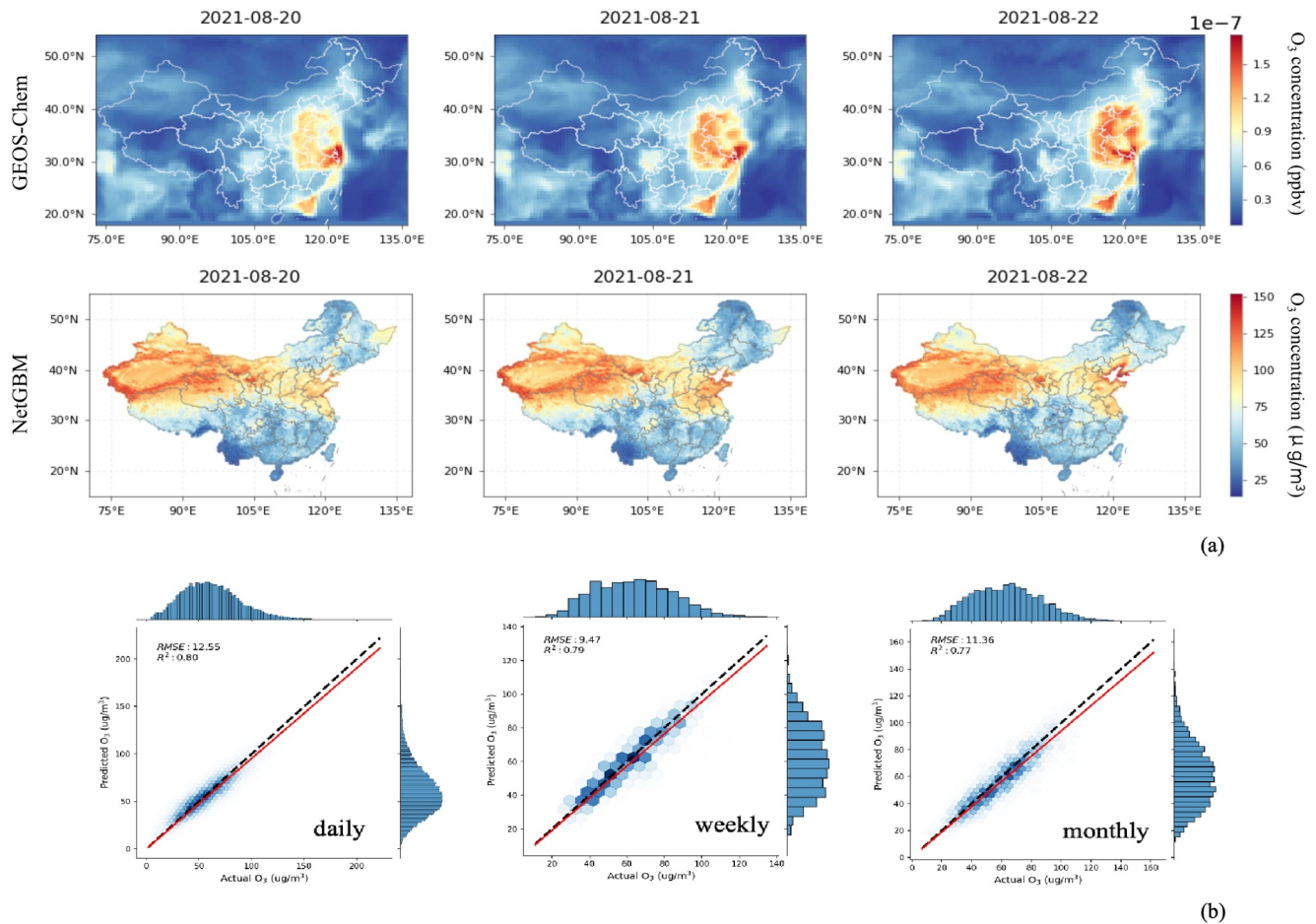


Figure 3. Prediction Results. (a) Comparison of the spatial distribution of O₃ across China predicted by GEOS-Chem and NetGBM from 20 to 22 August 2021; (b) Scatterplots comparing observed and predicted O₃ concentrations for (i) daily, (ii) weekly and (iii) monthly data.

and RMSE values were found to be 0.80 and 8.60 $\mu\text{g m}^{-3}$, respectively. This augmentation in performance is further accentuated in the context of monthly O₃ concentration estimations, spanning the temporal interval from 2019 to 2021 and derived from a sample of 15,993 observations. Herein, the R² and RMSE values exhibit additional improvement, quantified at 0.82 and 10.20 $\mu\text{g m}^{-3}$, respectively. The validation results can be found in Figure S5 in Supporting Information S1. Evidently, the model's predictive proficiency advances with the expansion of the temporal framework. This phenomenon may be attributed to the emergence of discernible patterns and trends within the O₃ concentration data as the temporal unit lengthens, facilitating the model's adept capture and accurate prognostication of these underlying patterns.

The evaluation of performance is depicted in Figure 3b, where scatterplots align observed measurements with O₃ predictions yielded by the model. Our comprehensive model demonstrates notable R² and RMSE values of 0.805 and 12.524 $\mu\text{g m}^{-3}$, respectively, for daily O₃ concentration predictions, consistent with validation outcomes. For weekly predictions, the model achieves an R² of 0.79 and an RMSE of 9.44 $\mu\text{g m}^{-3}$. Extending the prediction window to a monthly scale, the R² and RMSE are 0.77 and 11.27 $\mu\text{g m}^{-3}$, respectively. A comparison of our model's performance on daily ozone data with other models and previous works is presented in Table 1. For the USEPA-suggested metrics NMB and NME, NetGBM achieved −0.0152 and 0.1516, respectively, outperforming other models such as XGBoost and LightGBM. These results consistently align with the validation outcomes, affirming the robust predictive capabilities of NetGBM and its adeptness in mitigating overfitting to training data.

Figure 3a presents a comparison of the predicted ozone map utilizing our model against GEOS-Chem for the period of 20–22 August 2021. The fact that GEOS-Chem outputs include both tropospheric and stratospheric ozone and measure ozone at all vertical levels, while our methods utilized ozone readings from ground-level

Table 1
Comparison of Model Prediction Performance for Daily Predictions^a

Basic information			Performance				
Model (reference)	Study period	Resolution	R ²	RMSE	SMAPE	NMB	NME
MLR	2019.12–2021.11	0.25° × 0.25°	0.329	23.222	33.31	–	–
CAO ₃ (Mo et al., 2021)	2017.01–2017.12	0.1° × 0.1°	0.35	25.77	42.06	–	–
kNN	2019.12–2021.11	0.25° × 0.25°	0.644	16.915	23.38	–	–
Random Forest Regression (Zhan et al., 2018)	2015.01–2015.12	0.1° × 0.1°	0.69	26.00	–	–	–
Random Forest Regression*	2019.12–2021.11	0.25° × 0.25°	0.732	14.671	20.532	–	–
SVM			Null	Null	Null	Null	Null
XGBoost (R. Liu, Ma, et al., 2020)	2013.01–2018.12	0.1° × 0.1°	0.64	–	–	–	–
XGBoost*	2019.12–2021.11	0.25° × 0.25°	0.755	14.036	19.357	–0.0303	0.1770
LightGBM			0.782	13.226	18.607	–0.0370	0.1667
NN			0.759	13.93	19.768	–0.0140	0.2564
meteorological sensitivity			0.507	19.187	–	–0.0293	0.2304
NetGBM			0.805	12.524	17.663	–0.0152	0.1516

^aWe employed the identical Random Forest Regression and XGBoost methods following the approach detailed in references (Zhan et al., 2018) and (R. Liu, Ma, et al., 2020) respectively. Our experimentation involved applying these methods to both the data sets used in the mentioned references and our specific data set noted by *. All the studies utilize the same ozone data sourced from the national monitoring stations.

monitors, which provide at or near-surface ozone concentrations, leads to noticeable differences in the predicted ozone distributions. It is also worth noting that the outputs from GEOS-Chem are usually in units of ppbv (parts per billion by volume, volume mixing ratio). In the actual atmosphere, the total column ozone concentration is typically expressed in Dobson Units (DU). One DU represents a total ozone column amount equivalent to a 0.01 mm thick layer of pure ozone at standard temperature and pressure. The global average total column ozone concentration is approximately 300 DU, but it varies significantly by region and season. In contrast, ground-level ozone monitoring stations measure ozone in units typically in $\mu\text{g m}^{-3}$ (micrograms per cubic meter) or ppb (parts per billion, volume concentration). These measurements reflect ozone levels near the human activity layer and differ from the total column ozone concentration. For the evaluation, we searched for the closest ground-based simulated data corresponding to each monitoring station. The simulated data represents the Volume Mixing Ratio, which is the ratio of the volume of a particular substance (in our case, ozone) to the volume of the entire mixture. Due to the differences in units and monitoring heights, it is not possible to calculate metrics such as RMSE and MAE. Therefore, we used the Pearson correlation coefficient to calculate the correlation between the simulated data and the monitoring data. GEOS-Chem obtained a Pearson correlation coefficient of 0.36, showing its weakness in measuring surface-level ozone concentrations, while NetGBM achieved a Pearson correlation coefficient of 0.89, demonstrating that our model outperforms GEOS-Chem in predicting ozone distributions at or near the surface level, which is more critical for assessing the impact of ozone on public health.

To facilitate a comprehensive comparative analysis, our model is juxtaposed against established methodologies previously employed for estimating O₃ concentrations. This meticulous evaluation culminates in synthesizing findings, concisely presented in Tables 1 and 2. Notably, NetGBM distinguishes itself by showcasing significantly enhanced performance compared to its counterparts. Importantly, the majority of these models exhibit daily prediction accuracies with R² below 0.80. For instance, the CAMS ozone (CAO₃) prediction, with an accuracy as low as 0.35 (Mo et al., 2021), introduces pronounced biases into the estimates of O₃ concentrations. Moreover, the intricate nature of the daily data processed by the Support Vector Machine (SVM) presents challenges, often leading to execution timeouts due to the substantial computational complexity involved. Recent research endeavors concerning ozone distributions across China have been relatively limited in scope, with a predominant focus on specific regions or major cities. In our investigation, we discovered that Zhan et al. (2018), R. Liu, Ma, et al. (2020), and T. Liu, Wang, et al. (2020) applied national models based on the random forest and XGBoost techniques, yielding daily R² values of 0.69 and 0.64, respectively (T. Liu, Wang, et al., 2020; Zhan et al., 2018). In our work, we attempted to apply those same methods to our data set, elevating the prediction

Table 2
Comparison of Model Prediction Performance for Different Timescales

Model	Prediction								
	Daily			Weekly			Monthly		
Model (reference)	R ²	RMSE	SMAPE	R ²	RMSE	SMAPE	R ²	RMSE	SMAPE
MLR	0.329	23.222	33.31	0.386	16.352	22.253	0.338	19.081	26.544
CAO ₃ (Mo et al., 2021)	0.35	25.77	42.06	—	—	—	—	—	—
kNN	0.644	16.915	23.38	0.733	10.772	13.68	0.698	12.887	16.925
RF (Zhan et al., 2018)	0.69	26.00	—	—	—	—	—	—	—
RF*	0.732	14.671	20.532	0.777	9.859	12.198	0.77	11.227	14.326
SVM	Null	Null	Null	0.734	10.752	14.435	0.694	12.96	17.924
XGBoost (R. Liu, Ma, et al., 2020)	0.64	—	—	—	—	—	0.60–0.87	—	—
XGBoost*	0.755	14.036	19.357	0.787	9.64	11.424	0.768	11.294	14.074
LightGBM	0.782	13.226	18.607	0.772	9.969	12.696	0.748	11.757	15.291
NN	0.759	13.93	19.768	0.77	9.857	12.17	0.728	12.22	16.439
NetGBM	0.805	12.524	17.663	0.796	9.436	11.536	0.77	11.267	14.815

performances for both models to 0.732 and 0.75, respectively. While our research elevated prediction performance to R² values surpassing 0.70, these models were still outperformed by our comprehensive model.

It is also noteworthy that when applying the same model parameters to the identical data set, utilizing individual models LightGBM and NN, the resultant daily R² values consistently fell below those achieved by our integrated model. We also conducted a meteorological variation validation by testing the sensitivity of meteorological factors toward our model, with results shown in Table 1 as meteorological sensitivity. Without meteorological factors, the model presented a predicted R² as 0.507, and −0.0293 and 0.2304 for NMB and NME, respectively, showing that meteorological data are essential to our model's performance. This observation underscores the substantial enhancement achieved through the synergistic fusion of these two methodologies within NetGBM, thereby accentuating the markedly superior performance demonstrated by our integrated approach.

3.2. Spatial-Temporal Distributions

The research, until now, has been based on the whole country. For a more detailed evaluation of the model, we separated China into several different regions and looked at the model's prediction performance in those areas respectively. We take daily O₃ concentration data separately from six representative regions, that is:

- NEP—the Northeast Plain of China, including Jilin, Changchun and Heilongjiang, with bitterly cold winters, fertile land and moderate rainfall. It faces considerable air pollution issues primarily due to industrial activities, and coal burning (X. Li, Hu, et al., 2019).
- NCP—the North China Plain, located in the middle-north of China, centered at Beijing, China's most significant plain combined with a semi-humid monsoon climate and cold, dry winters. This area also suffers from severe air pollution, with high concentrations of airborne pollutants, particulates, and heavy metals linked to heavy industry, vehicle emissions, and coal combustion (K. Li et al., 2021).
- YRD—the Yangtze River Delta at the southeast coast of China, including Jiangsu, Zhejiang and Shanghai, which is full of water channels and lakes and has ample rainfall. It is one of China's most economically developed regions, while one of the others is SCP. This region's primary air pollution source is vehicle emissions and industrial outputs, but with some related policies, its air quality has improved in recent years (L. Li et al., 2020; T. Liu, Wang, et al., 2020).
- SCP—the South China Plain on the south coast of China, including Guangdong, Guangxi, Hainan, Hong Kong and Macao. It is intersected by rivers and canals, combined with a humid subtropical climate, plentiful rainfall and steamy summers. The air pollution situation in this area is similar to that of YRD, with some significant improvements in these years (M. Hu et al., 2021).

Table 3
Supporting Static and Dynamic Features Used in This Study

Data category	Data type	Data description	Resolution
Geographic and Land Use factors	Elevation (m)	The elevation of the pointed locations	0.25° × 0.25°
	Grass Land (%)	Percentage of grassland coverage	
	Wood Land (%)	Percentage of woodland coverage	
	Water Coverage (%)	Percentage of water coverage.	
	Ocean Coverage (%)	Percentage of ocean coverage	
	Urban-Rural Ratio (%)	The ratio between the area of the urban areas to that of rural areas within the grid	
	Cultivated Land (%)	The percentage of areas are actively used for agricultural purposes	
	Unused Land (%)	The percentage of areas that are not currently utilized for any specific purpose, including agriculture, residential, industrial, or recreational activities	
Meteorology and Atmospheric factors	Relative humidity (%)	The meteorological data were collected from fixed monitoring stations across China and subsequently averaged for each grid cell, as illustrated in Figure 2c	0.25° × 0.25° × 1d
	Specific humidity (%)		
	Pressure (kPa)		
	Precipitate Water (mm)		
	Temperature (°C)		
	AOD		
	Wind direction ([0°, 360°])		
	Wind speed (m/s)		
Pollutant information	O ₃ concentration	The concentration of ozone, collected from fixed monitoring stations across China and averaged into each grid	0.25° × 0.25° × 1d

Note. The study period is from 1st December 2019 to 30th November 2021.

- TiP—the Tibetan Plateau in China. This region is immensely rugged, with high altitudes and diverse land-forms like mountains and gorges. With lower population density, this region has better air quality than the populated plains and coastal regions in the East. Air pollution in this area can also come from desert dust (L. Zhang et al., 2022).

The predictive performance can be visually discerned in Figure 4a, along with the distribution of crucial features across the country. The results demonstrate the satisfactory performance of our model nationwide, with an average R^2 surpassing 0.70 across all regions. Notably, variations in performance are evident across distinct geographical areas. The most notable performance is observed in the NCP region, characterized by high population density and predominantly dry and clean atmospheric conditions. Here, the R^2 values predominantly exceed 0.85, resulting in an impressive average of 0.858. Following closely, the YRD region exhibits a commendable predictive capability, with most R^2 values surpassing 0.80 and an average of 0.823. Similarly, the NEP region showcases a noteworthy precision in predicting O₃ concentrations, with an average R^2 value of 0.811. The predictive prowess in the SCP region, characterized by relatively high humidity and frequent overcast weather, manifests within a range of 0.50–0.80, primarily exceeding 0.70. In contrast, the TiP region exhibits comparatively lower prediction accuracy, with average R^2 values of 0.742. Notably, some stations in the TiP region yield R^2 values lower than 0.5. These distinctive results highlight the model's augmented performance in regions defined by clear atmospheres and dry weather conditions. The comparatively diminished prediction accuracy noted within the TiP region might be ascribed to the scarce quantity of monitoring stations and the diverse sources of ozone. Predominantly, the ozone pollution in this area is mainly a result of stratospheric intrusion rather than human activities, as suggested in a recent study (J. Yang et al., 2022). In areas with limited monitoring infrastructure, the NetGBM model demonstrated its reliability, with predictions falling within 10% of actual measurements 85% of the time. This highlights the potential of machine learning to provide accurate ozone estimates in regions with sparse monitoring networks (Requia et al., 2020).

It is essential to acknowledge that the foundation of our research is established on data spanning the years 2019–2021. This timeframe assumes particular significance due to the emergence of the COVID-19 pandemic in China,

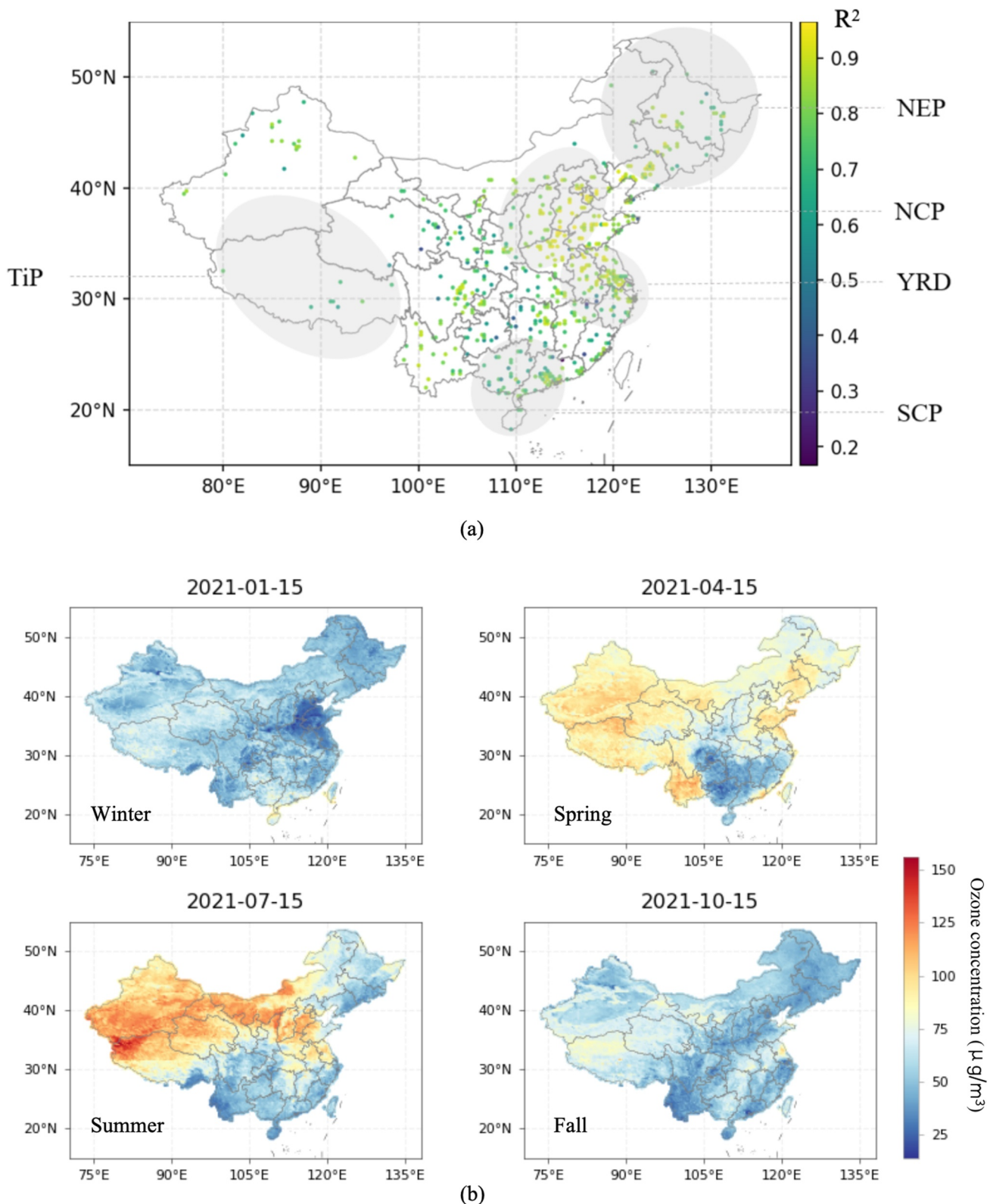


Figure 4. The comparison of observed and predicted O_3 on a daily scale throughout China, as well as some of the major features across the country: (a) The scattering of R^2 across different regions of China; (b) Spatial distributions of the seasonal example of the predicted daily Ozone concentrations for China during (i) winter, (ii) spring, (iii) summer, and (iv) fall.

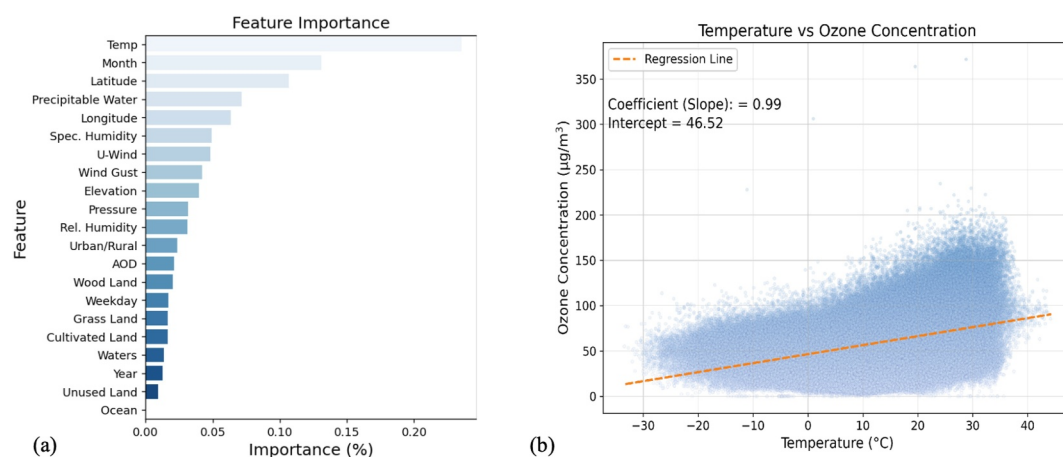


Figure 5. Visualization of feature importance results. (a) The feature importance ranks for all features used in this experiment; (b) trend of ozone concentrations change with temperature.

a pivotal factor with the potential to profoundly influence air quality dynamics by implementing local lockdown measures (Y. Zhu et al., 2020). Stringent industrial lockdowns were instituted by various governmental bodies to contain the virus's transmission and ensure public safety. These measures engendered a discernible reduction in industrial activities, resulting in a substantial decline in air pollution levels. This unique context is likely to have exerted an impact on the predictive outcomes of our model. A notable illustration of this phenomenon is observed in the province of Jilin, situated within the NEP. Subsequent research has corroborated a marked reduction in pollution levels during these periods of enforced inactivity. Despite these exceptional circumstances, NetGBM has exhibited commendable validation performance within the NEP. The observed results remain robust and reliable, as evidenced by an R^2 value exceeding 0.80. By incorporating historical data, the model successfully captured the paradoxical increase in ozone levels by 5% during the COVID-19 lockdown periods, despite a 15% reduction in NO_2 concentrations due to decreased vehicular traffic, which aligns with previous findings (S. Zhu et al., 2021). Moreover, the model's high-resolution predictions revealed that urban areas can experience 20% higher ozone concentrations compared to rural regions, potentially leading to a 10% increase in respiratory-related illnesses among the urban population, as supported by prior studies (Hong et al., 2019). Therefore, despite the intricate and atypical variables introduced by the pandemic, our model has demonstrated resilience and validation.

3.2.1. Feature Analysis

Numerous factors, including population density, pollution levels, land use, and meteorological conditions, contribute to regional disparities in air quality. Predicting O_3 concentrations can be particularly challenging under conditions of low atmospheric visibility. Incorporating meteorological features and AOD data enhanced the predictive capabilities of the NetGBM model. The results indicate that temperature is the most influential factor in determining ozone concentrations, as shown in Figure 5a. Additionally, the features “Month” and “Latitude” significantly impact ozone concentrations due to spatial-temporal variations in temperature. Specifically, the model demonstrated that a 1°C increase in temperature corresponds to an approximately 1 ppb increase in ozone concentration, as shown in Figure 5b, which is consistent with previous research (Zhong et al., 2021). It underscores the need for temperature-adjusted ozone action plans, especially in urban heat islands. Furthermore, incorporating AOD data improved the model's accuracy by 5%, suggesting that on days with high AOD, ozone levels were, on average, 8 ppb lower due to the scattering and absorption of solar radiation by aerosols. These findings are consistent with previous studies (Abdi-Oskouei et al., 2020; Q. Liu et al., 2019).

Notably, meteorological features exert significant influence, indicating that sites characterized by clear atmospheres yield more accurate results due to the precision of NO_x and VOCs data. This suggests that locations with higher humidity levels may exhibit improved predictive performance, consistent with findings from prior studies (Abdullah et al., 2019). The performance of the model is also intimately tied to pollution levels. Specifically, in areas with high concentrations of O_3 pollution such as NCP and NEP, there is a noticeable enhancement in the predictions of O_3 concentrations. This increase in performance becomes more significant in the summer when the

pollution intensity escalates (K. Li et al., 2021). Furthermore, regions with higher population densities like YRD and NCP have also demonstrated superior model performance, possibly due to the sufficient investment in air monitoring equipment, facilitating a richer data set for the model's consumption (L. Li et al., 2020). In addition to these factors, disparities in land use and land cover contribute to variations in predictive outcomes. Notably, the contrast in land use patterns between Eastern and Western China is striking. The Eastern region is characterized by a higher concentration of urban areas associated with economic advancement and substantial population densities, although green spaces such as forests and grasslands are limited. Conversely, Western regions, with a prevalence of rural areas, feature sparser populations spread over extensive plateaus. Predictive performance in these areas is notably lower due to fewer inhabitants, significant forest cover, and extensive undeveloped land. Additional information on the static and dynamic features can be found in Section Materials and Methods.

3.2.2. Predictions at Region-Specific Training Set

Another crucial consideration is recognizing that, until now, the model's training has utilized data from across the entire nation, whereas its predictive capabilities have been applied in distinct and separate regional contexts. Previous research has demonstrated that prediction performance can be significantly impacted when the model is deployed on different scales. Further refinement could potentially be achieved by training the model on localized data, leveraging the strong correlation between spatial attributes and O_3 concentrations (J. Yang & Zhao, 2023). In light of this, we have selected two regions, NCP and YRD, as examples, to train the NetGBM model using data from within these regions and applied the model to these specific local features.

The resulting outcomes yielded average R^2 values of 0.892 and 0.879 for NCP and YRD, respectively. In the context of a 10-fold validation, the model's performances in both geographical regions were reflected in RMSE values exceeding 12.00, precisely 12.59 for NCP and 12.24 for YRD. The respective SMAPE values were 14.75 for NCP and 15.59 for YRD, while MAE values were 8.77 for NCP and 8.95 for YRD. Notably, a substantial number of the predicted R^2 outcomes surpassed 0.90, signifying a high degree of predictive accuracy for both regional-based models.

The outcomes of the NCP region are elucidated in Figures 6a and 6b, wherein the robust correlations between predicted and observed O_3 concentrations are prominently evident. Similarly, the corresponding figure for the YRD region underscores the model's commendable performance as depicted in Figures 6c–6f. A comparative analysis of R^2 values using YRD as an example is shown in Figures 6e and 6f. Notably, these results surpassed those obtained from the national model, proving that our model performs more effectively in location-specific contexts. However, it is important to note that while performance improvements are evident, the extent of variance may not be substantial enough to assert that training on localized features yields significant enhancements definitively. In conclusion, NetGBM has demonstrated robustness in estimating O_3 concentrations at both national and region-specific scales, with the latter yielding improved performance.

3.3. Ozone Exposure and Health Risk Assessment

To explore the potential of this method in assessing the health impacts of ozone, we calculated ozone exposure volumes and the mortality attributable to ozone exposure, as described in Section 2.5. Applying our data to the Chinese population in 2021, we estimated both the ozone exposure inhalation metrics and the mortality rates for five major causes of death (COPD, IHD, LC, LRI, and stroke) using GEMM (Burnett et al., 2018). In this study, we focused on the highest ozone exposure concentrations throughout the year and applied these data to the population aged 40–50 as an example to demonstrate the applicability of our method in health risk assessment.

Figure 7a presents the GEMM predictions for each of the five CODs, with stroke specifically analyzed for the 45-year age group. Figure 7b illustrates the spatial distribution of ozone-attributable mortality for these causes of death, and Table S1 in Supporting Information S1 details locations with the highest mortality rates, including their corresponding longitudes and latitudes. IHD and stroke contributed most significantly to the total ozone-attributable mortality, jointly causing around 200 million deaths in 2021, which is more than 70% of the total deaths. The mortality hotspots predominantly appear in densely populated urban areas such as NCP, YRD, and SCP. This information is valuable for ozone management in China. For instance, according to Table S1 in Supporting Information S1, the highest ozone-attributed IHD mortality in 2021 was located in the grid centered at latitude 31.25 and longitude 121.5, precisely corresponding to Hongkou District, Shanghai. Similarly, Beijing,

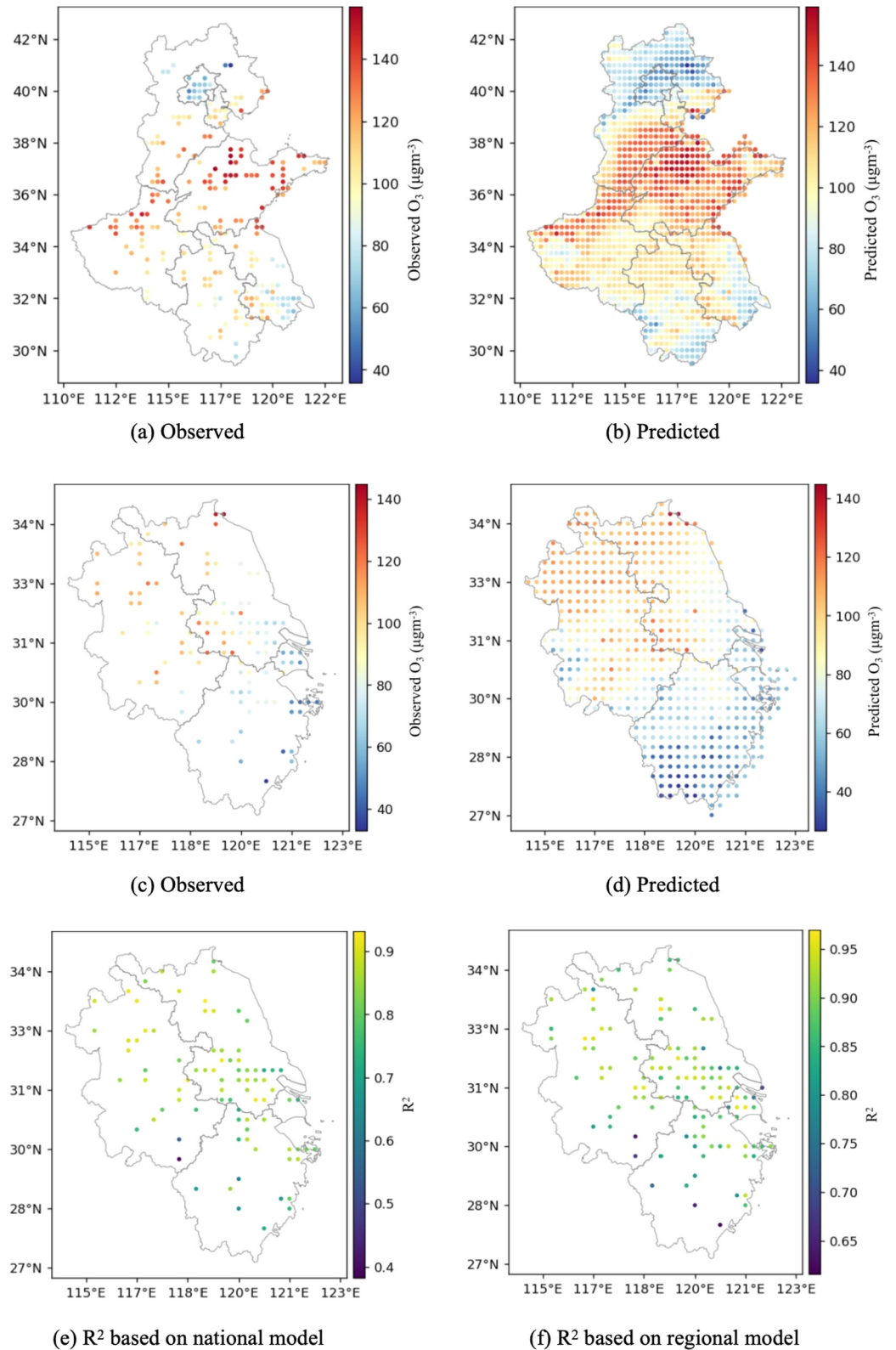


Figure 6. Performance evaluation in NCP and YRD. (a) Observed O_3 in NCP; (b) predicted O_3 in NCP; (c) Observed O_3 in YRD; (d) predicted O_3 in YRD; The prediction R^2 on YRD based on (e) national data model and (f) regional data model.

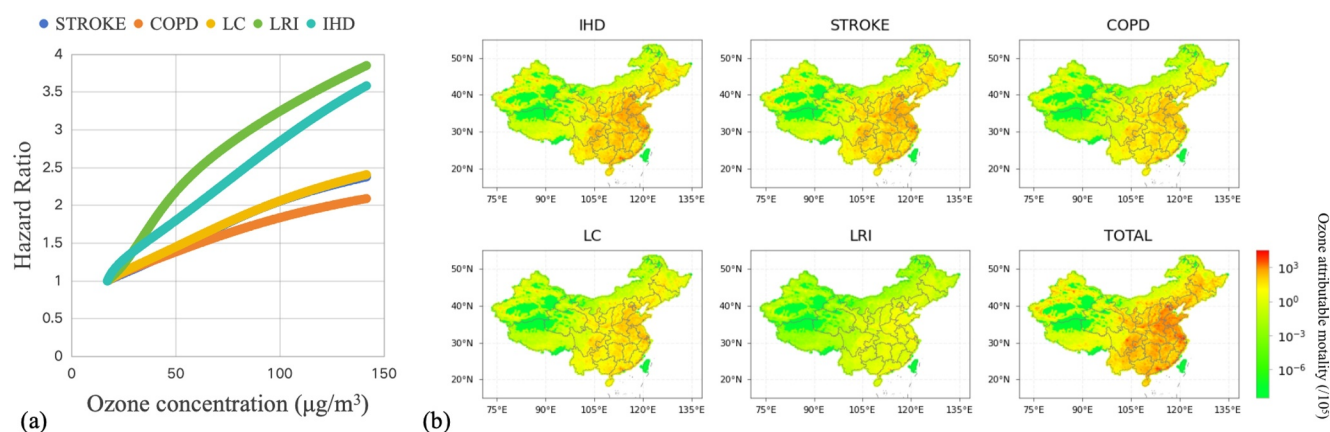


Figure 7. GEMM Results: (a) Hazard ratios for 5-CODs as determined by GEMM. (b) Spatial distribution of ozone-attributable mortality categorized by cause of death.

Guangdong, and Jiangsu all appeared in the top five grids with the highest mortality. On the other hand, although TiP has high ozone concentrations, the mortality is nearly 0 due to the low population density. The increase in ozone concentrations will have a larger impact on the relative risk of LRI and IHD. Such detailed information can assist policymakers in prioritizing areas for ozone management. For a visualization example, we selected ozone inhalation volumes and ozone concentrations from 20 to 25 August 2021, as shown in Figures S6 and S7 in Supporting Information S1. This approach allows for the derivation of ozone inhalation volume data by date, enabling applications such as high ozone inhalation warnings and targeted policy decisions based on identified hotspots. Our study has established a potential method for high-resolution health assessments and mortality predictions across China, contributing significantly to health management in related fields.

It is important to note that the experiment conducted in this research utilized annual average data and parameters for a specific age group to provide an example of mortality; although the relative ratios of the results hold, the actual values may be biased due to this data selection process. Nevertheless, the major contribution of this study is that it provides a foundation for future work to conduct more in-depth and thorough health risk assessments. Additionally, the method has the potential to be applied in air pollution management, which will be the focus of our future research endeavors.

4. Conclusion

In conclusion, our study addresses a critical challenge: the fine-scaled and robust prediction of ozone concentrations across diverse spatial and temporal domains. We also analyzed the associated ozone exposure and health risks across China. Through the innovative fusion of advanced machine learning methodologies, we have developed a pioneering NetGBM model that effectively models ozone levels on various timescales. The model has demonstrated satisfactory and robust performance across daily, weekly, and monthly predictions, with R^2 coefficients of 0.83, 0.77, and 0.79, respectively. To the best of our knowledge, compared with previous works on ozone predictions across China, our approach achieves superior and more robust performance by incorporating geographical and land use information, meteorological data, different timescales, and AOD data. The model's performance is further enhanced in regions with high population density and pollution levels.

Compared to previous studies, our research provides a more recent and detailed analysis of ozone distribution utilizing high spatial-temporal resolution ($0.25^\circ \times 0.25^\circ$, daily) data across China that includes the pandemic period in 2021. Our findings indicate a significant increase in national ozone levels in 2021, with temperature and spatial-temporal scales being the primary factors influencing ozone concentrations in China. Precipitation and humidity also exhibit notable impacts. In our health assessment, we applied our predicted ozone concentrations to inhalation analysis and GEMM, concluding that although ozone concentrations are elevated in both western regions and the eastern coast, ozone-attributed mortality is more severe in developed areas such as NCP, YRD, and SCP, particularly in large urban centers. Therefore, effective ozone management in the post-pandemic period, especially during hot seasons and in densely populated and polluted regions, is crucial.

Given the performance of our model in providing timely and fine-scaled ozone predictions and exposure distributions, we propose packaging our model into a product that continuously delivers real-time information across China. This will enable timely provision of ozone exposure and health risk data, facilitate pollution warnings, and provide a basis for further analysis of ozone emissions in China, such as source tracking and public health management. Nonetheless, a noteworthy limitation of NetGBM stems from its reliance solely on ground-based monitoring stations for daily data collection. This predisposes the model to potential inaccuracies in assessing O_3 concentrations, possibly inducing performance disparities across regions characterized by varying monitoring station densities. To address this limitation, future research avenues should encompass a fusion of data from mobile and satellite monitors, bolstering the precision of O_3 predictions. Additionally, there is a need to analyze hourly concentrations or even finer temporal resolutions, refining the model's temporal resolution and insights. Furthermore, it is worth noting that both the Net and LightGBM have limited the model's interpretability. Future research endeavors should focus on exploring models with higher interpretability to gain a better understanding of how these models operate in predicting pollutants across China.

5. Materials

5.1. Data Gridded Map

Our study utilized daily O_3 concentration data collected from 1,733 air quality monitoring stations throughout China from 1 December 2019, to 30 November 2021, as depicted in Figure 1a. This data served as the foundation for creating a national network. We divided the national map into grids measuring $0.25^\circ \times 0.25^\circ$, allocating the mean Ozone concentration for each grid based on the average readings of the stations within the respective grid. The same approach was employed to gather meteorological attributes from weather stations within each grid, with daily averages calculated for every grid. On the other hand, static urban aspects such as green coverage and land use were documented and segmented into discrete grids, which will be further explained in the next section. Ultimately, the nation was partitioned into a cumulative total of 747 grids, resulting in 530,370 observations over 710 days. This count was obtained by multiplying 747 grids by 710 days.

Upon completing data allocation, quality assurance steps were undertaken, involving spot verification and data normalization processes to eliminate discernible anomalies and ensure requisite data quality. Over 90% of the data were retained after the quality assurance process, resulting in an aggregate of 482,727 observation samples earmarked for our investigation. When examining larger timescales, such as weekly and monthly durations, the accrued daily data was consolidated and categorized into weekly and monthly intervals, and the model was consequently run on each respective timescale. An arbitrary splitting pattern was adopted for training and testing data sets to validate the gridded network's ability to portray real-time Ozone conditions accurately. Specifically, 90% of the grids across the country were randomly selected as the training set, and the subsequent data was utilized as the testing set. Similarly, for the regional model, a randomized 90% of the local grids were chosen as the training set. It is in this latter section that observations were juxtaposed with predicted data to evaluate the model's performance.

5.2. Meteorological and Spatial Land-Use Features

The data set under consideration comprises 21 distinct features for each observation, as shown in Table 3. In our feature importance comparison, those features are also thoughtfully categorized into eight distinct groups, each serving a specific analytical purpose. Within the context of each observation, it encompasses (a) Location features, encompassing parameters such as latitude and longitude. (b) Terrain attributes, including elevation information and indicating whether the observation is over an oceanic region. (c) Land Use features are also involved, encompassing parameters related to cultivated land, urban-rural ratios, vegetation cover (including wood and grass), water coverage, and areas of unused land, as visually depicted in Figure S3 in Supporting Information S1. Furthermore, the data set incorporates meteorological features, specifically (d) Atmospheric Parameters, encompassing critical metrics such as temperature, pressure, and precipitable water content. (e) Humidity-related attributes are also considered, encompassing specific and relative humidity values. (f) Wind patterns are also investigated, involving U-wind components and wind gust details. (g) Time Cycle attributes further enhance the data set, providing information about the month, year, and specific day of the week for each observation. Additionally, the data set includes (h) AOD as a distinctive feature type. Illustrated examples of these dynamic meteorological features can be found in Figure S4 in Supporting Information S1, which presents a seasonal comparison of some significant features between summer and winter, utilizing data from February 1st

and August 1st, respectively. We applied a feature importance analysis toward the mentioned features, where the relative importance of each feature is succinctly ranked as shown in Figure 5a. These rigorously vetted features serve as a pivotal cornerstone in our subsequent experimental process. Observations characterized by static and dynamic attributes are designated as the input vector V and V_t , respectively, which is subsequently concatenated to form the composite input vector employed by NetGBM.

5.3. Neural Network and LightGBM Parameters

In this research, we employ both a neural network and a LightGBM model to perform predictive analyses. The neural network architecture includes an input layer, four hidden layers, and an output layer, all interconnected through fully connected layers. The first hidden layer comprises 1,680 neurons, the second hidden layer contains 2,560 neurons, the third hidden layer has 128 neurons, and the fourth hidden layer includes 64 neurons. The output layer is configured to produce a single-dimensional output. Each hidden layer is followed by a batch normalization layer to enhance training stability and a ReLU activation function to introduce non-linearity. Additionally, a dropout layer with a dropout probability of 0.2 is incorporated to mitigate overfitting by randomly deactivating a fraction of neurons during training. The forward propagation sequence is as follows: The input passes through the first fully connected layer, generating feature1. The output then undergoes batch normalization and ReLU activation, followed by dropout. This process is iteratively repeated until the output traverses the fifth fully connected layer, resulting in the main output.

For the LightGBM model, the parameters are meticulously chosen to optimize performance. The model uses the Gradient Boosting Decision Tree (GBDT) as the boosting type and applies regression to predict continuous values. The model's performance is evaluated using RMSE metric. To control model complexity, 1,500 leaves are set, and the maximum depth is limited to 20, which dictates the maximum number of leaves and the depth of each tree, respectively. The learning rate is configured at 0.05 to regulate the step size during each iteration. To enhance generalization, the feature fraction is set to 0.9, and the bagging fraction is set to 0.8, ensuring that a random subset of features and data samples is selected for each iteration. The bagging frequency is set to five, indicating the frequency of bagging operations.

Data Availability Statement

We have uploaded the code and results generated and used during this study, which are available in the permanent online repository (C. Ma & Ran, 2024).

Acknowledgments

The work is supported by the Smart Society Lab at Hong Kong Baptist University.

References

- Abdi-Oskouei, M., Carmichael, G., Christiansen, M., Ferrada, G., Rooszitalab, B., Sobhani, N., et al. (2020). Sensitivity of meteorological skill to selection of WRF-Chem physical parameterizations and impact on ozone prediction during the lake Michigan ozone study (LMOS). *Journal of Geophysical Research: Atmospheres*, 125(5), e2019JD031971. <https://doi.org/10.1029/2019jd031971>
- Abdullah, S., Nasir, N. H. A., Ismail, M., Ahmed, A. N., & Jarkoni, M. N. K. (2019). Development of ozone prediction model in urban area. *International Journal of Innovative Technology and Exploring Engineering*, 8(10), 2263–2267. <https://doi.org/10.35940/ijitee.j1127.0881019>
- Allu, S. K., Srinivasan, S., Maddala, R. K., Reddy, A., & Anupoju, G. R. (2020). Seasonal ground level ozone prediction using multiple linear regression (MLR) model. *Modeling Earth Systems and Environment*, 6(4), 1981–1989. <https://doi.org/10.1007/s40808-020-00810-0>
- AlOmar, M. K., Hameed, M. M., & AlSaadi, M. A. (2020). Multi hours ahead prediction of surface ozone gas concentration: Robust artificial intelligence approach. *Atmospheric Pollution Research*, 11(9), 1572–1587. <https://doi.org/10.1016/j.apr.2020.06.024>
- Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., III, et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences of the United States of America*, 115(38), 9592–9597. <https://doi.org/10.1073/pnas.1803222115>
- Chen, Z., Liu, J., Cheng, X., Yang, M., & Shu, L. (2023). Stratospheric influences on surface ozone increase during the COVID-19 lockdown over northern China. *npj Climate and Atmospheric Science*, 6(1), 76. <https://doi.org/10.1038/s41612-023-00406-2>
- Cheng, Y., Huang, X.-F., Peng, Y., Tang, M.-X., Zhu, B., Xia, S.-Y., & He, L.-Y. (2023). A novel machine learning method for evaluating the impact of emission sources on ozone formation. *Environmental Pollution*, 316, 120685. <https://doi.org/10.1016/j.envpol.2022.120685>
- Duan, X. (2015). *Highlights of the Chinese exposure factors handbook*. Academic Press.
- Fu, T.-M., & Tian, H. (2019). Climate change penalty to ozone air quality: Review of current understandings and knowledge gaps. *Current Pollution Reports*, 5(3), 159–171. <https://doi.org/10.1007/s40726-019-00115-6>
- Henze, D. K., Hakami, A., & Seinfeld, J. H. (2007). Development of the adjoint of GEOS-Chem. *Atmospheric Chemistry and Physics*, 7(9), 2413–2433. <https://doi.org/10.5194/acp-7-2413-2007>
- Hong, C., Zhang, Q., Zhang, Y., Davis, S. J., Tong, D., Zheng, Y., et al. (2019). Impacts of climate change on future air quality and human health in China. *Proceedings of the National Academy of Sciences of the United States of America*, 116(35), 17193–17200. <https://doi.org/10.1073/pnas.1812881116>
- Hu, J., Chen, J., Ying, Q., & Zhang, H. (2016). One-year simulation of ozone and particulate matter in China using WRF/CMAQ modeling system. *Atmospheric Chemistry and Physics*, 16(16), 10333–10350. <https://doi.org/10.5194/acp-16-10333-2016>

- Hu, M., Wang, Y., Wang, S., Jiao, M., Huang, G., & Xia, B. (2021). Spatial-temporal heterogeneity of air pollution and its relationship with meteorological factors in the Pearl River delta, China. *Atmospheric Environment*, 254, 118415. <https://doi.org/10.1016/j.atmosenv.2021.118415>
- Iglesias-Gonzalez, S., Huertas-Bolanos, M. E., Hernandez-Paniagua, I. Y., & Mendoza, A. (2020). Explicit modeling of meteorological explanatory variables in short-term forecasting of maximum ozone concentrations via a multiple regression time series framework. *Atmosphere*, 11(12), 1304. <https://doi.org/10.3390/atmos11121304>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- Le, T., Wang, Y., Liu, L., Yang, J., Yung, Y. L., Li, G., & Seinfeld, J. H. (2020). Unexpected air pollution with marked emission reductions during the COVID-19 outbreak in China. *Science*, 369(6504), 702–706. <https://doi.org/10.1126/science.abb7431>
- Li, J., Nagashima, T., Kong, L., Ge, B., Yamaji, K., Fu, J. S., et al. (2019). Model evaluation and intercomparison of surface-level ozone and relevant species in east Asia in the context of mics-Asia phase III—part 1: Overview. *Atmospheric Chemistry and Physics*, 19(20), 12993–13015. <https://doi.org/10.5194/acp-19-12993-2019>
- Li, K., Jacob, D. J., Liao, H., Qiu, Y., Shen, L., Zhai, S., et al. (2021). Ozone pollution in the north China plain spreading into the late-winter haze season. *Proceedings of the National Academy of Sciences of the United States of America*, 118(10), e2015797118. <https://doi.org/10.1073/pnas.2015797118>
- Li, L., Li, Q., Huang, L., Wang, Q., Zhu, A., Xu, J., et al. (2020). Air quality changes during the COVID-19 lockdown over the Yangtze River delta region: An insight into the impact of human activity pattern changes on air pollution variation. *Science of the Total Environment*, 732, 139282. <https://doi.org/10.1016/j.scitotenv.2020.139282>
- Li, X., Hu, X., Shi, S., Shen, L., Luan, L., & Ma, Y. (2019). Spatiotemporal variations and regional transport of air pollutants in two urban agglomerations in northeast China plain. *Chinese Geographical Science*, 29(6), 917–933. <https://doi.org/10.1007/s11769-019-1081-8>
- Liu, Q., Liu, T., Chen, Y., Xu, J., Gao, W., Zhang, H., & Yao, Y. (2019). Effects of aerosols on the surface ozone generation via a study of the interaction of ozone and its precursors during the summer in Shanghai, China. *Science of the Total Environment*, 675, 235–246. <https://doi.org/10.1016/j.scitotenv.2019.04.121>
- Liu, R., Ma, Z., Liu, Y., Shao, Y., Zhao, W., & Bi, J. (2020). Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. *Environment International*, 142, 105823. <https://doi.org/10.1016/j.envint.2020.105823>
- Liu, T., Wang, X., Hu, J., Wang, Q., An, J., Gong, K., et al. (2020). Driving forces of changes in air quality during the COVID-19 lockdown period in the Yangtze River delta region, China. *Environmental Science and Technology Letters*, 7(11), 779–786. <https://doi.org/10.1021/acs.estlett.0c00511>
- Lu, X., Zhang, L., Wang, X., Gao, M., Li, K., Zhang, Y., et al. (2020). Rapid increases in warm-season surface ozone and resulting health impact in China since 2013. *Environmental Science and Technology Letters*, 7(4), 240–247. <https://doi.org/10.1021/acs.estlett.0c00171>
- Luecken, D., Yarwood, G., & Hutzell, W. (2019). Multipollutant modeling of ozone, reactive nitrogen and HAPs across the continental US with CMAQ-CB6. *Atmospheric Environment*, 201, 62–72. <https://doi.org/10.1016/j.atmosenv.2018.11.060>
- Ma, C., & Ran, M. (2024). First release: Machine learning-driven ozone exposure analysis. *Zenodo*. <https://doi.org/10.5281/zenodo.13756158>
- Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., He, M. Z., et al. (2021). Random forest model based fine scale spatiotemporal O₃ trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. *Environmental Pollution*, 276, 116635. <https://doi.org/10.1016/j.envpol.2021.116635>
- Marty, M. A., Blaisdell, R. J., Broadwin, R., Hill, M., Shimer, D., & Jenkins, M. (2002). Distribution of daily breathing rates for use in California's air toxics hot spots program risk assessments. *Human and Ecological Risk Assessment: An International Journal*, 8(7), 1723–1737. <https://doi.org/10.1080/20028091057574>
- Mo, Y., Li, Q., Karimian, H., Zhang, S., Kong, X., Fang, S., & Tang, B. (2021). Daily spatiotemporal prediction of surface ozone at the national level in China: An improvement of CAMS ozone product. *Atmospheric Pollution Research*, 12(1), 391–402. <https://doi.org/10.1016/j.apr.2020.09.020>
- Ren, X., Mi, Z., & Georgopoulos, P. G. (2020). Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environment International*, 142, 105827. <https://doi.org/10.1016/j.envint.2020.105827>
- Requia, W. J., Di, Q., Silvern, R., Kelly, J. T., Koutrakis, P., Mickley, L. J., et al. (2020). An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. *Environmental Science & Technology*, 54(18), 11037–11047. <https://doi.org/10.1021/acs.est.0c01791>
- Song, J., Fan, H., Gao, M., Xu, Y., Ran, M., Liu, X., & Guo, Y. (2022). Toward high-performance map-recovery of air pollution using machine learning. *ACS ES&T Engineering*, 3(1), 73–85. <https://doi.org/10.1021/acsestengg.2c00248>
- Song, J., Han, K., & Stettler, M. E. (2020). Deep-maps: Machine-learning-based mobile air pollution sensing. *IEEE Internet of Things Journal*, 8(9), 7649–7660. <https://doi.org/10.1109/jiot.2020.3041047>
- Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., de Hoogh, K., et al. (2020). A random forest approach to estimate daily particulate matter, nitrogen dioxide, and ozone at fine spatial resolution in Sweden. *Atmosphere*, 11(3), 239. <https://doi.org/10.3390/atmos11030239>
- Tie, X., Geng, F., Guenther, A., Cao, J., Greenberg, J., Zhang, R., et al. (2013). Megacity impacts on regional ozone formation: Observations and WRF-Chem modeling for the mirage-shanghai field campaign. *Atmospheric Chemistry and Physics*, 13(11), 5655–5669. <https://doi.org/10.5194/acp-13-5655-2013>
- Wang, P., Wang, P., Chen, K., Du, J., & Zhang, H. (2022). Ground-level ozone simulation using ensemble WRF/Chem predictions over the southeast United States. *Chemosphere*, 287, 132428. <https://doi.org/10.1016/j.chemosphere.2021.132428>
- Wei, J., Li, Z., Pinker, R. T., Wang, J., Sun, L., Xue, W., et al. (2021). Himawari-8-derived diurnal variations in ground-level pm 2.5 pollution across China using the fast space-time light gradient boosting machine (LightGBM). *Atmospheric Chemistry and Physics*, 21(10), 7863–7880. <https://doi.org/10.5194/acp-21-7863-2021>
- Yang, G., Liu, Y., & Li, X. (2020). Spatiotemporal distribution of ground-level ozone in China at a city level. *Scientific Reports*, 10(1), 7229. <https://doi.org/10.1038/s41598-020-64111-3>
- Yang, J., Wang, K., Lin, M., Yin, X., & Kang, S. (2022). Not biomass burning but stratospheric intrusion dominating tropospheric ozone over the Tibetan Plateau. *Proceedings of the National Academy of Sciences of the United States of America*, 119(38), e2211002119. <https://doi.org/10.1073/pnas.2211002119>
- Yang, J., & Zhao, Y. (2023). Performance and application of air quality models on ozone simulation in China—a review. *Atmospheric Environment*, 293, 119446. <https://doi.org/10.1016/j.atmosenv.2022.119446>
- Zhan, Y., Luo, Y., Deng, X., Grieneisen, M. L., Zhang, M., & Di, B. (2018). Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. *Environmental Pollution*, 233, 464–473. <https://doi.org/10.1016/j.envpol.2017.10.029>

- Zhang, L., Zhang, H., & Xu, E. (2022). Information entropy and elasticity analysis of the land use structure change influencing eco-environmental quality in Qinghai-Tibet Plateau from 1990 to 2015. *Environmental Science and Pollution Research*, 29(13), 18348–18364. <https://doi.org/10.1007/s11356-021-17978-2>
- Zhang, T., Yue, X., Unger, N., Feng, Z., Zheng, B., Li, T., et al. (2021). Modeling the joint impacts of ozone and aerosols on crop yields in China: An air pollution policy scenario analysis. *Atmospheric Environment*, 247, 118216. <https://doi.org/10.1016/j.atmosenv.2021.118216>
- Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., et al. (2019). A predictive data feature exploration-based air quality prediction approach. *IEEE Access*, 7, 30732–30743. <https://doi.org/10.1109/access.2019.2897754>
- Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z., & Huang, L. (2020). A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Transactions*, 100, 210–220. <https://doi.org/10.1016/j.isatra.2019.11.023>
- Zheng, B., Zhang, Q., Geng, G., Chen, C., Shi, Q., Cui, M., et al. (2021). Changes in China's anthropogenic emissions and air quality during the COVID-19 pandemic in 2020. *Earth System Science Data*, 13(6), 2895–2907. <https://doi.org/10.5194/essd-13-2895-2021>
- Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., et al. (2021). Robust prediction of hourly PM_{2.5} from meteorological data using LightGBM. *National Science Review*, 8(10), nwaa307. <https://doi.org/10.1093/nsr/nwaa307>
- Zhu, S., Poetscher, J., Shen, J., Wang, S., Wang, P., & Zhang, H. (2021). Comprehensive insights into O₃ changes during the COVID-19 from O₃ formation regime and atmospheric oxidation capacity. *Geophysical Research Letters*, 48(10), e2021GL093668. <https://doi.org/10.1029/2021gl093668>
- Zhu, Y., Xie, J., Huang, F., & Cao, L. (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *Science of the Total Environment*, 727, 138704. <https://doi.org/10.1016/j.scitotenv.2020.138704>