atmosphere

MDPI

Article

High-Spatiotemporal-Resolution Estimation of Ground-Level Ozone in China Based on Machine Learning

Jiahuan Chen [1], Heng Dong [1,2], Zili Zhang [3,4], Bingqian Quan [3] and Lan Luo [5,*]

[1] School of Resources and Environment Engineering, Wuhan University of Technology, Wuhan 430070, China; 302352@whut.edu.cn (J.C.); simondong@whut.edu.cn (H.D.)
[2] Zhejiang Spatiotemporal Sophon Bigdata Co., Ltd., Ningbo 315101, China
[3] Ecological Environment Monitoring Center of Zhejiang, Hangzhou 310012, China; zhangzili@zjemc.org.cn (Z.Z.); quanbingqian@zjemc.org.cn (B.Q.)
[4] Zhejiang Key Laboratory of Ecological Environment Monitoring, Early Warning and Quality Control Research, Hangzhou 310012, China
[5] Zhejiang Key Laboratory of Ecological and Environmental Big Data (2022P10005), Zhejiang Ecological and Environmental Monitoring Center, Hangzhou 310012, China
* Correspondence: luolan@zjemc.org.cn

Abstract: High concentrations of ground-level ozone (O_3) pose a significant threat to human health. Obtaining high-spatiotemporal-resolution information about ground-level O_3 is of paramount importance for O_3 pollution control. However, the current monitoring methods have a lot of limitations. Ground-based monitoring falls short in providing extensive coverage, and remote sensing based on satellites is constrained by specific spectral bands, lacking sensitivity to ground-level O_3. To address this issue, we combined brightness temperature data from the Himawari-8 satellite with meteorological data and ground-based station data to train four machine learning models to obtain high-spatiotemporal-resolution information about ground-level O_3, including Categorical Boosting (CatBoost), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and Random Forest (RF). Among these, the CatBoost model exhibited superior performance, achieving a ten-fold cross-validation R^2 of 0.8534, an RMSE of 17.735 μg/m^3, and an MAE of 12.6594 μg/m^3. Furthermore, all the selected feature variables in our study positively influenced the model. Subsequently, we employed the CatBoost model to estimate averaged hourly ground-level O_3 concentrations at a 2 km resolution. The estimation results indicate a close relationship between ground-level O_3 concentrations and human activities and solar radiation.

Keywords: ground-level ozone; high-spatiotemporal-resolution; machine learning

check for updates

Citation: Chen, J.; Dong, H.; Zhang, Z.; Quan, B.; Luo, L. High-Spatiotemporal-Resolution Estimation of Ground-Level Ozone in China Based on Machine Learning. *Atmosphere* **2024**, *15*, 34. https://doi.org/10.3390/atmos15010034

Academic Editor: Alexandros Papayannis

Received: 14 November 2023
Revised: 8 December 2023
Accepted: 22 December 2023
Published: 27 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As a trace gas in the atmosphere, 90% of ozone (O_3) is dispersed in the stratosphere between 10 and 50 km from the ground, and the remaining 10% of atmospheric O_3 is distributed in the troposphere below 10 km from the ground [1,2]. O_3 in the stratosphere protects Earth's organisms from the damaging effects of ultraviolet radiation [3]. In contrast, excessively high ground-level O_3 concentrations not only emit pungent odors but also irritate the human respiratory system, causing damage to lung cells and posing significant risks to human health [4–6]. According to the World Health Organization (WHO), humans are subjected to life and health threats when exposed to maximum 8 h average O_3 concentrations exceeding the recommended threshold of ≥100 μg/m^3 [7]. However, according to the Ministry of Ecology and Environment of the People's Republic of China, the annual average O_3 concentrations in 339 Chinese cities have all surpassed the WHO's recommended threshold of 100 μg/m^3 [8]. This indicates that ground-level O_3 pollution poses a significant hazard to the health of Chinese residents and the ecological environment. O_3 pollution urgently needs to be addressed. Notably, the high-spatiotemporal-resolution estimation of ground-level O_3 is a crucial step in addressing O_3 pollution issues [9,10].

Currently, ground-based station monitoring and satellite sensor monitoring are the two main methods for monitoring the spatiotemporal distribution of ground-level O_3. As of 2021, China has established 2024 national monitoring stations for trace gases. However, these stations are mainly concentrated in provincial capitals and central cities, resulting in an uneven spatial distribution and the incapacity to provide high-resolution, continuous, and extensive spatiotemporal O_3 distribution information. In the short term, it is difficult for China to establish a dense and extensive monitoring network for trace gases. Therefore, relying solely on ground-based station monitoring methods is inadequate for meeting China's current requirements for addressing O_3 pollution [11].

Compared to ground-based station monitoring, satellite remote sensing monitoring is not constrained by time, climate, or geographical limitations, facilitating large-scale synchronous observations and providing extensive spatial coverage [12–14]. For instance, in 2015, the Japan Aerospace Exploration Agency (JAXA) successfully launched the Himawari-8 satellite, which has a 10 min observation frequency [15]. The satellite is equipped with the Advanced Himawari Imagers (AHI) sensor, which can provide brightness temperature (BT) data products with a spatial resolution of 2 km in multiple thermal infrared (TIR) bands. Based on TIR bands, the Infrared Atmospheric Sounding Interferometer (IASI) can directly monitor the vertical O_3 profile staring from the ground, which has a correlation coefficient of 0.85 in the validation comparing to ground-based measurements [16,17]. In addition, the BT at the TIR bands show a positive correlation with solar radiation intensity [18]. Thus, BT products from AHI are presently being used in various research studies to produce high-spatiotemporal-resolution O_3 distribution information [19,20]. However, owing to the specific portion of the electromagnetic spectrum used by satellite sensors, the current satellite instruments have limited sensitivity to ground-level O_3. Relying solely on remote sensing observations also makes it challenging to achieve the precise monitoring of ground-level O_3 [21,22].

For the estimation of O_3 concentrations, numerous methodologies have been widely implemented. There are many ways to estimate O_3 concentrations, including frameworks of chemical transport models (CTMs) and statistical models. A CTM typically consists of four main components: physical transport, pollutant emissions, dispersion, and chemical transformation. Depending on various input parameters, the model can integrate and process pollutant concentrations for a specific period, providing the average pollutant concentration during that interval [23]. Some scholars have utilized CTMs to investigate the spatiotemporal distribution of ground-level O_3 concentrations, such as the global 3-D CTM from the Goddard Earth Observing System (GEOS-chem) [24] and the Copernicus Atmosphere Monitoring Service (CAMS) [25]. CTMs comprehensively consider various physical, chemical, and dynamical atmospheric processes. They have precise physical and chemical meanings and possess strong interpretability. However, due to limited knowledge and input data, the fine-scale predictions of atmospheric chemistry models may deviate considerably from the actual results, and their ability to predict the spatial and temporal distribution of high-resolution ozone concentrations may need to be improved [26].

For statistical methods, initially, spatial interpolation methods, such as inverse distance weighting, which are relatively straightforward and cost-effective, were employed [27–29]. Then, traditional statistical models, which have evolved from linear regression to more complex methods that can incorporate many geographic features and satellite-derived data, such as geographic-temporal weighted regression models and land-use regression models, have emerged to estimate O_3 information. For instance, Kerckhoffs [30] devised a land-use regression model that centers on summer average O_3 concentrations and annual average O_3 concentrations as the primary exposure variables. This model effectively accounts for 71% of the spatial variability in summer average O_3 concentrations.

In recent years, the use of machine learning models based on multi-source data to estimate O_3 concentrations has become a prominent area of research. Felder [31] constructed a neural network O_3 inversion system. This system utilized automatic feature selection and automatic architecture search to reduce the training time by approximately two orders

of magnitude, thereby rendering the O_3 concentrations inversion system more stable. Zhan [32] combined meteorological data, elevation data, emission inventories, normalized difference vegetation indices (NDVI), land use data, and road density data to estimate the daily maximum 8 h average O_3 concentrations in the region of China in 2015 with the random forest model (RF). The results of the cross-validation indicated an R^2 of 0.69 and an RMSE of 26 $\mu g/m^3$. Li [33] initially used the RF model to patch in missing total O_3 column data over Hainan Island, China. They then employed the eXtreme Gradient Boosting algorithm to estimate ground-level O_3 concentrations over Hainan Island based on the total O_3 column and other estimated parameters. The model obtained an R^2 of 0.59 and an RMSE of 6.36 $\mu g/m^3$. Li [34] employed a gradient boosting regression tree algorithm, incorporating ground-level O_3 concentration data, MODIS NDVI data, weather research and forecasting (WRF) meteorological data, and population data. They used a backward variable selection method to train the model with the best feature variables, resulting in a distribution of high-spatiotemporal-resolution ground-level O_3 concentrations. The model obtained an R^2 of 0.89 and an RMSE of 4.75 $\mu g/m^3$ in cross-validation. These findings indicate that machine learning models exhibit exceptional performance when it comes to estimating O_3 concentrations. However, in existing studies, either the spatial or temporal resolution is always coarse (e.g., $0.75^\circ \times 0.75^\circ$ with three-hourly measurements in CAMS), which will be challenging to provide effective support for the precise control of ozone in China.

In order to obtain high-spatiotemporal-resolution information on ground-level O_3 concentrations, we integrated data from the AHI, ground-based stations, and ERA5-Land (meteorological data), and we contrasted prominent machine learning models, which have become popular in recent years for their fast training speeds, high efficiency, and accurate predictions, including Categorical Boosting (CatBoost), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and RF, to determine the best-performing model. Finally, we estimated the average hourly spatiotemporal distribution of O_3 over one week by implementing the optimal model and feature variables. This research intends to provide a scientific basis and methodological support for the control and prevention of O_3 pollution.

2. Materials and Methods

2.1. Study Area

China is situated in eastern Asia on the west coast of the Pacific Ocean, with latitudes extending from 4° N to 53° N and longitudes from 73° E to 135° E. China has a land area of 9.6 million km^2 and a complex topography, which is characterized by a topography of high in the west and low in the east, with mountainous terrain dominating the west and plains and hills dominating the east. The topography of China decreases in a sequence of one, two, and three steps from the Tibetan Plateau to the north and east. Since the full-disk scanning area of the Himawari-8 satellite is 80° E–160° W, 60° S–60° N [15], it unable to thoroughly cover the Xinjiang Autonomous Region and Tibet Autonomous Region, so these two provinces were excluded from the study area. In recent years, China has made significant efforts to address environmental pollution issues, including the establishment of a network of in situ stations to monitor air pollutants such as ground-level trace gases. We selected the 1935 in situ stations that covered the study area by 2021 (see Figure 1). These stations are primarily concentrated in densely populated areas such as provincial capitals and central cities, and there are not enough monitoring stations in suburban counties and townships.

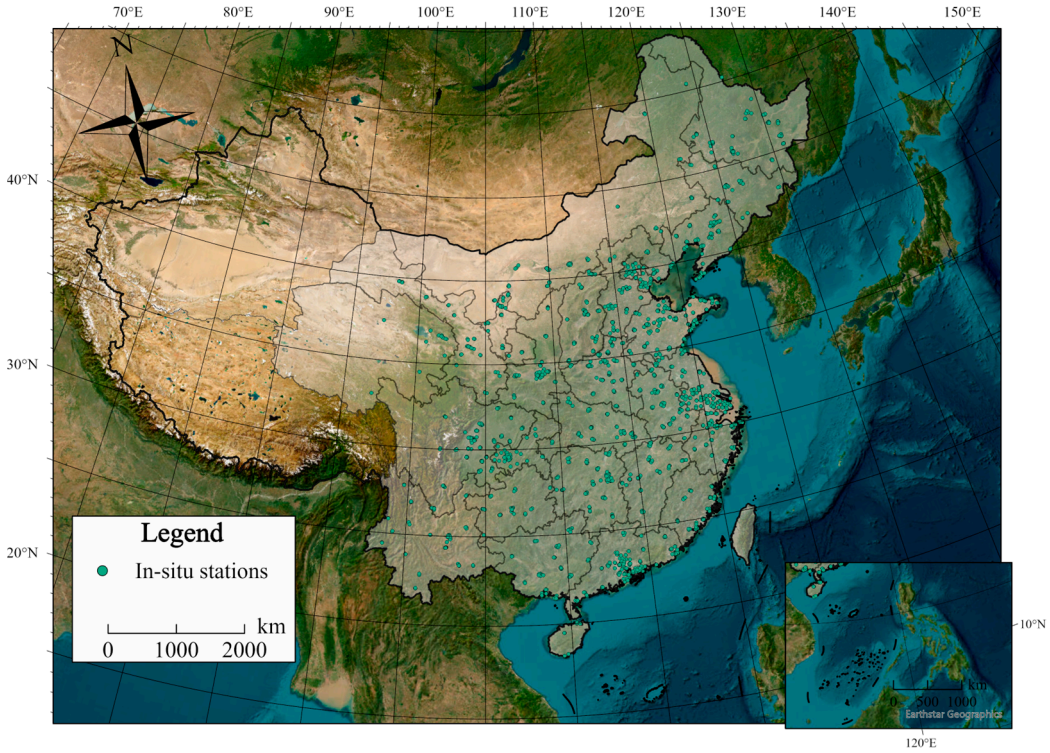

Figure 1. The distribution of in situ stations in the study area. The study area is delineated by a light grey shading. The base-map is the global imagery provided by Earthstar Geographics.

2.2. Datasets and Preprocessing

According to previous research, the high-spatiotemporal-resolution distribution of O_3 can be determined using infrared radiation with a wavelength of 9.6 μm measured by geostationary satellites. In this paper, the AHI BT data products were selected as the primary input parameters and were combined with meteorological data from ERA5-Land as the auxiliary input parameters (Table 1). In previous research, in order to estimate the O_3 concentrations near the ground surface, some scholars accounted for the influence of anthropogenic and topographic factors and frequently analyzed popular data, land use data, and road network data. Nevertheless, according to the research findings of Li [35], Zhao [36], and others [37,38], in the estimation of O_3 concentrations at a high spatial-temporal resolution, the characteristic variables such as terrain, surface cover, and road networks are slow or nearly unchanged on the time scale, and the dispersion of the characteristics is poor, which has a negative impact on the training of the model and the estimation performance, etc. Therefore, this research did not consider feature variables such as DEM (digital elevation model), NDVI, and road density.

Table 1. Data list used in the study area and related information.

Data Source	Data Name	Spatial Resolution	Time Resolution
JAXA	AHI BT data (band 10–band 16 except band 11)	0.02° × 0.02°	10 min
ERA5-Land	2 m temperature (T2M) 2 m dewpoint temperature (D2M) The top-net solar radiation (TSR) The boundary layer height (BLH) The surface latent heat flux (SLHF)	0.25° × 0.25°	1 h
CNEMC	Ground-level station data		1 h

2.2.1. AHI Bright Temperature Data

In our study, the wavelength of 9.6 μm (the absorption peak of O_3) in the AHI BT data product (band 12) was used as the model's primary input parameter. In addition, band 8 (wavelength of 6.2 μm)–band 10 (wavelength of 7.3 μm) and band 13 (wavelength of 10.4 μm)–band 16 (wavelength 13.3 μm) were selected as the model auxiliary parameters according to Lee et al. In addition, band 11 (wavelength of 8.6 μm) was excluded from the estimation of ground-level O_3 concentrations because it is particularly susceptible to desert emissivity fluctuations. All of the AHI BT data were at a resolution of 10 min $0.02^\circ \times 0.02^\circ$.

2.2.2. Meteorological Data

Taking into account the significant impact of meteorological conditions on ozone formation [39,40], we utilized meteorological data as the auxiliary information for the model. The meteorological data in our study were derived from ERA5-Land (a reanalysis dataset) provided by the European Center for Medium-Range Weather Forecasts (ECMWF). Based on the laws of physics, ECMWF produced the reanalysis dataset by combining model data with observations from across the world. ERA5-Land dataset can provide hourly meteorological products at a resolution of hourly $0.25^\circ \times 0.25^\circ$. Considering that the effect of meteorological conditions on O_3 is not instantaneous, we used the ERA5-Land portion of the meteorological data from 08:00–17:00 (UTC + 8) from 1 June 2021 to 31 December 2021, including the 2 m temperature (T2M), the 2 m dewpoint temperature (D2M), the top-net solar radiation (TSR), the boundary layer height (BLH), and the surface latent heat flux (SLHF).

2.2.3. Ground-Based Station Data

The China National Environmental Monitoring Centre (CNEMC) provided the hourly ground-level O_3 concentration data from 09:00–18:00 (UTC + 8) for the period from 1 June 2021 to 31 December 2021 for 1935 in situ stations in the study area. In accordance with the HJ818-2018 standard, CNEMC employs ultraviolet dual-beam detection technology by the ozone standard reference photometer to measure ground-level O_3 concentrations. In our study, we used the station data provided by CNEMC as the true O_3 concentrations to train our models.

2.2.4. Data Preprocessing

To assure the consistent spatial resolution of our input data set, we applied the IDW to resample meteorological data from ERA5-Land to 2-km. Due to the incapacity of AHI to mitigate atmospheric scattering and cloud interference, AHI data cannot accurately capture ground-level information in cloud-covered regions. To address this issue, we eradicate cloud-contaminated pixels using the daytime cloud property product (L2CLP) provided by JAXA. In the L2CLP, each pixel is classified according to the cloud classification standards of the World Meteorological Organization (WMO), where pixels with "CLTYPE = 0" represent those not covered by clouds. In the process of cloud removal, we therefore overlayed the daytime cloud attribute product with the brightness temperature data, retaining pixels with "CLTYPE = 0" and removing those with other values for "CLTYPE."

We then performed a temporal and spatial alignment of ground-based station data with AHI and ERA5-Land data. In terms of time alignment, BT data are available at a temporal resolution of 10 min, whereas data from ground-based stations and ERA5-Land are hourly. To ensure temporal uniformity, we averaged the BT data within each hour, reducing it to an hourly temporal resolution. In addition, acknowledging that meteorological factors do not have an instantaneous effect on O_3 [41,42], we advanced the alignment of meteorological data with ground-based station data by one hour. For instance, the meteorological data for 8:00 on a particular day was matched with the station data for 9:00 on the same day. Regarding spatial alignment, we matched the ground-based station data with other datasets that fell within the same grid by extracting the attribute values to the grid.

2.3. Models

RF is an ensemble learning method based primarily on the construction of multiple decision trees for classification or regression tasks. Each decision tree is trained on a random subset of the data with random feature selection (using the bootstrap sampling method). The final forecast is determined by a vote or average of all trees. By integrating multiple decision trees, RF improves model performance by exhibiting strong resistance to noise and outliers.

Categorical Boosting (CatBoost) is a gradient boosting algorithm, uniquely characterized by its adoption of gradient boosting strategies to progressively refine prediction results. It automatically handles data encoding without the need for manual intervention, reducing the workload of feature engineering. Simultaneously, it mitigates the risk of overfitting.

eXtreme Gradient Boosting (XGBoost) is a highly optimized gradient boosting algorithm known for its ability to train multiple weak learners and then integrate them into a powerful model by optimizing the loss function. The capacity of XGBoost to manage massive datasets and intricate relationships is considerable. It employs regularization techniques to reduce model complexity and mitigate the risk of overfitting. Additionally, it facilitates parallel computation, which speeds up the training process.

Light Gradient Boosting Machine (LGBM) is a gradient boosting algorithm based on histograms that is renowned for its extraordinary performance and memory efficiency. It accelerates data partitioning by constructing histograms. LGBM employs a leaf-wise growth strategy as opposed to traditional depth-first strategies, allowing for faster training on large-scale datasets. Additionally, LGBM allows for the customization of loss functions and evaluation metrics.

In our study, we employed the above-mentioned four machine learning models to capture the nonlinear relationship between the input feature variables and ground-level O_3 concentrations.

2.4. Model Evaluation

We employed a ten-fold cross-validation (CV) method to evaluate the estimation accuracy of various machine learning models in both spatial and temporal dimensions. All matching grids were divided into ten subsets at random. The machine learning models were trained using nine subsets, while the remaining subset was used for validation. This procedure was carried out ten times. The estimation results were validated by three metrics: the coefficient of determination (R^2), the root mean square error (RMSE), and the mean absolute error (MAE).

3. Results and Discussion

In this research, a total of 1,070,869 samples were obtained after data preprocessing and sample selection (removal of outliers and zero values). Firstly, the complete dataset was randomly divided into a training dataset comprising 70% of the samples and a testing dataset comprising 30% to preliminarily train the models to adjust the model's hyperparameters. After the hyperparameter tuning, the models were evaluated based on the principle of ten cross-validation processes.

3.1. Feature Evaluation

Evaluating and selecting features is crucial for maximizing the performance of a model and improving the accuracy of predictions. In order to ascertain the positive contribution of the selected features to the models, we evaluate them from two different perspectives. One way we analyzed the relationship between the features and the target variable was by calculating the Pearson correlation coefficients (PCCs), as shown in Figure 2. The PCC figure clearly demonstrates the strong correlation between the meteorological parameters, including the T2M, BLH, TSR, and BT data from band 12 and band 13, and ground-level O_3 concentrations.

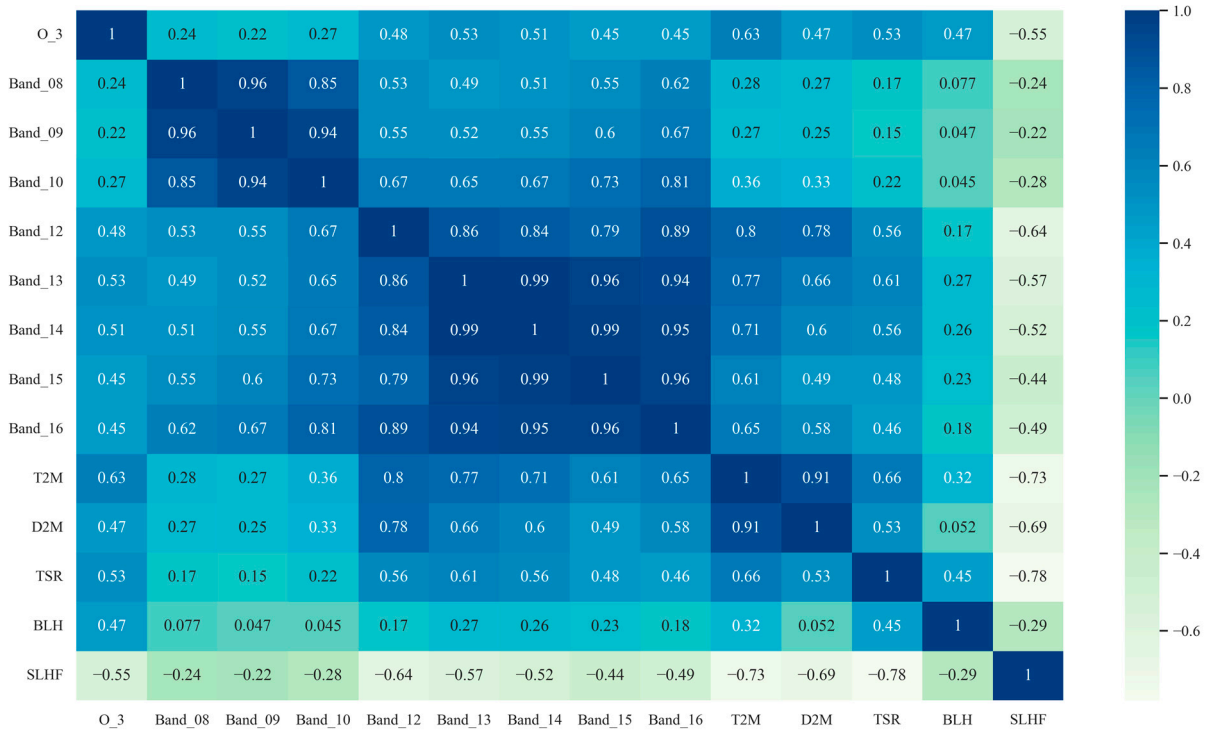

Figure 2. The Pearson correlation coefficients among various feature variables and correlations with the ground-level O_3 concentrations.

Then, we calculated importance coefficients for each feature with respect to the four machine learning models, as depicted in Figure 3. In each model, the meteorological factors T2M and BLH demonstrated significant importance, which was consistent with the computation of the PCC. The observed outcome can be attributed to the reduction in the BLH, resulting in the accumulation of O_3 precursors, namely nitrogen oxides (NOx) and volatile organic compounds (VOCs), in close proximity to the surface [43]. Moreover, T2M is essential in the chemical reaction that results in the synthesis of O_3 from the precursors NOx and VOC. The contribution of features was more evenly distributed in both the CatBoost and LGBM models, especially in the LGBM model. In the CatBoost model, the BT data of band 12, in addition to meteorological conditions, also made a substantial contribution. Regarding the XGBoost and RF models, the BT data largely functioned as a model correction component, making a smaller contribution to the model compared to T2M and BLH, which played more major roles.

Subsequently, we carried out a systematic process of feature reduction. This involved starting with the features that were determined to be the least essential based on their importance coefficients in various models. For each feature removal, we reported the model's validation metrics (R^2, RMSE, MAE) on the test dataset. The procedure is depicted in Figure 4. When features were eliminated one by one individually in the CatBoost, XGBoost, and RF models, the R^2 of the models generally decreased, while the RMSE and MAE typically increased. This suggests a progressive deterioration in the performance of the models. Nevertheless, the LGBM model exhibited a slight increase in R^2, accompanied by decreased RMSE and MAE values, upon the removal of the first feature. This indicates a moderate improvement in the model's performance. The reason for this could be that LGBM employs a leaf-wise growth approach as opposed to the conventional depth-first technique. This strategy prioritizes increasing the depth of trees rather than expanding all branches at each level. Eliminating one feature could potentially enhance tree segmentation, resulting in an improved performance of the model. To optimize the predictive performance, we eliminated the feature with the lowest importance scores for LGBM while keeping the features unchanged in the other models.

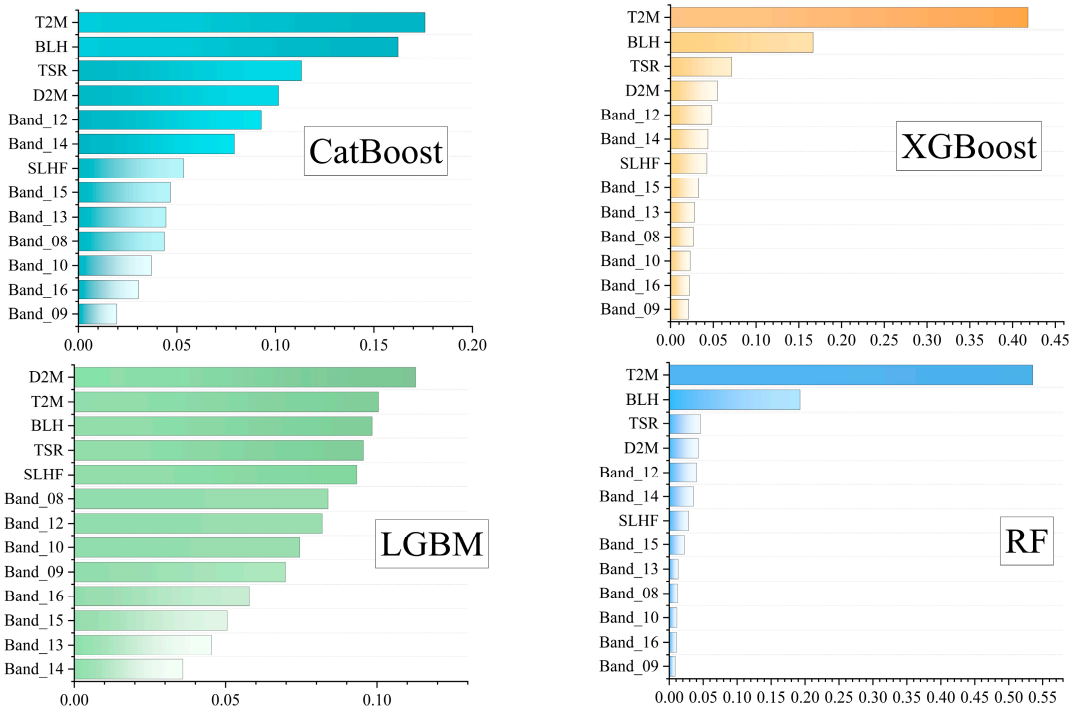

Figure 3. Importance coefficients of various features in each model.

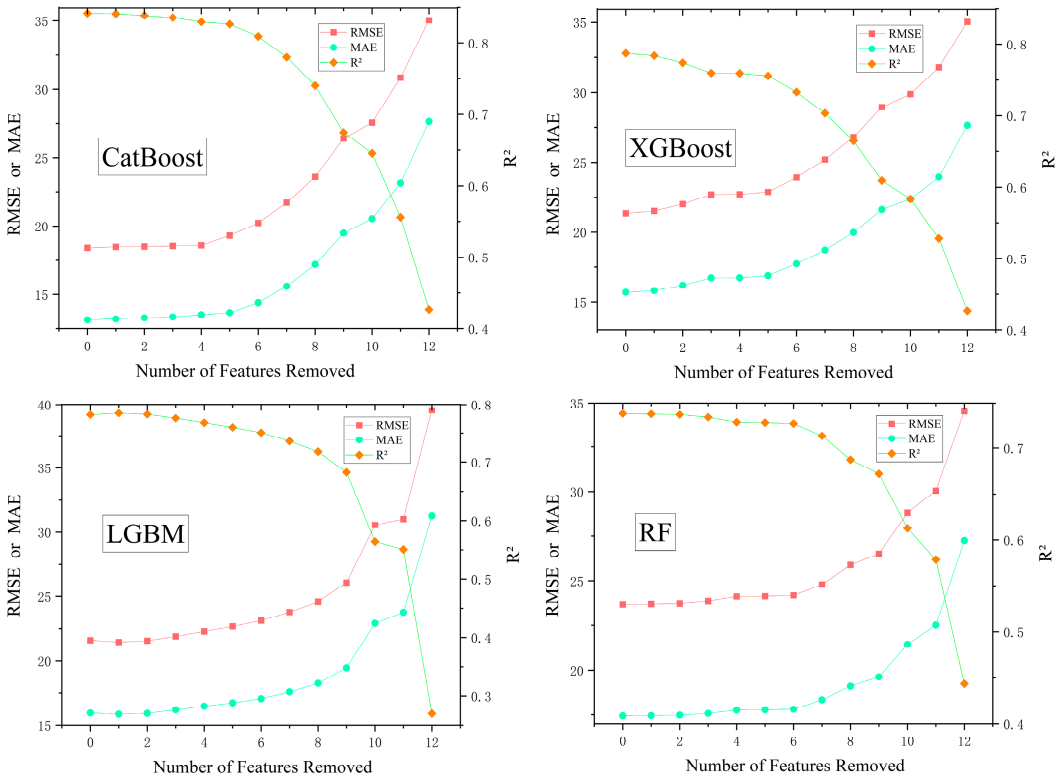

Figure 4. Performance of models with sequential feature reduction. The units of RMSE and MAE are $\mu g/m^3$.

3.2. Performance Analysis of Models

Based on the results discussed in Section 3.1, we trained the four models using the features that achieved the best predictive performance for each model. The CV performance

of each model is shown in Table 2 and Figure 5. The R^2 values for the CatBoost, XGBoost, LGBM, and RF models were 0.8534, 0.7947, 0.7872, and 0.7424, respectively. The RMSE values were 17.735 μg/m^3, 20.987 μg/m^3, 21.367 μg/m^3, and 23.510 μg/m^3, respectively. The MAE values were 12.6594 μg/m^3, 15.4337 μg/m^3, 15.8119 μg/m^3, and 17.3154 μg/m^3, respectively. From CatBoost to RF, the model's fitting performance rapidly diminished, and the errors for the target variable progressively increased.

Table 2. The validated metrics for each model.

Model Name	R^2	RMSE	MAE
CatBoost	0.8534	17.735	12.6594
XGBoost	0.7947	20.987	15.4337
LGBM	0.7872	21.367	15.8119
RF	0.7424	23.510	17.3154

The units of RMSE and MAE are μg/m^3.

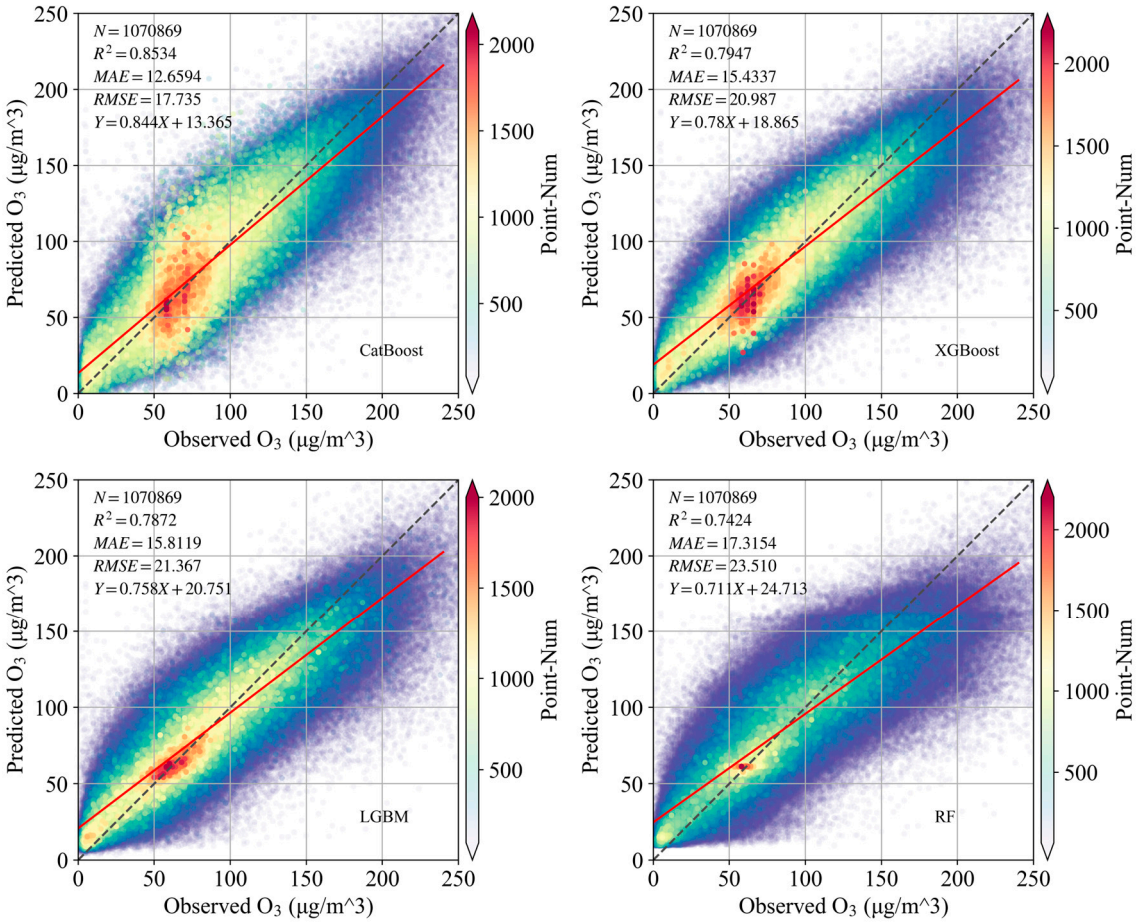

Figure 5. Density scatter diagrams between predicted O_3 concentrations and observed O_3 concentrations based on CV. The units of RMSE and MAE are μg/m^3.

3.3. Discussion of Spatiotemporal Distribution of O_3

We chose the week from 20 September to 26 September 2021, which had the lowest level of cloud contamination. We performed hourly assessments of ground-level O_3 concentrations from 09:00 to 18:00 (UTC + 8) for each day throughout that week and subsequently calculated their average. Figure 6 displays the results of the multi-day average estimation of ground-level O_3 concentrations for each hour. Following that, we conducted a statistical analysis on the estimated hourly ground-level O_3 concentrations, computing the mean and standard deviation, as shown in Figure 7.

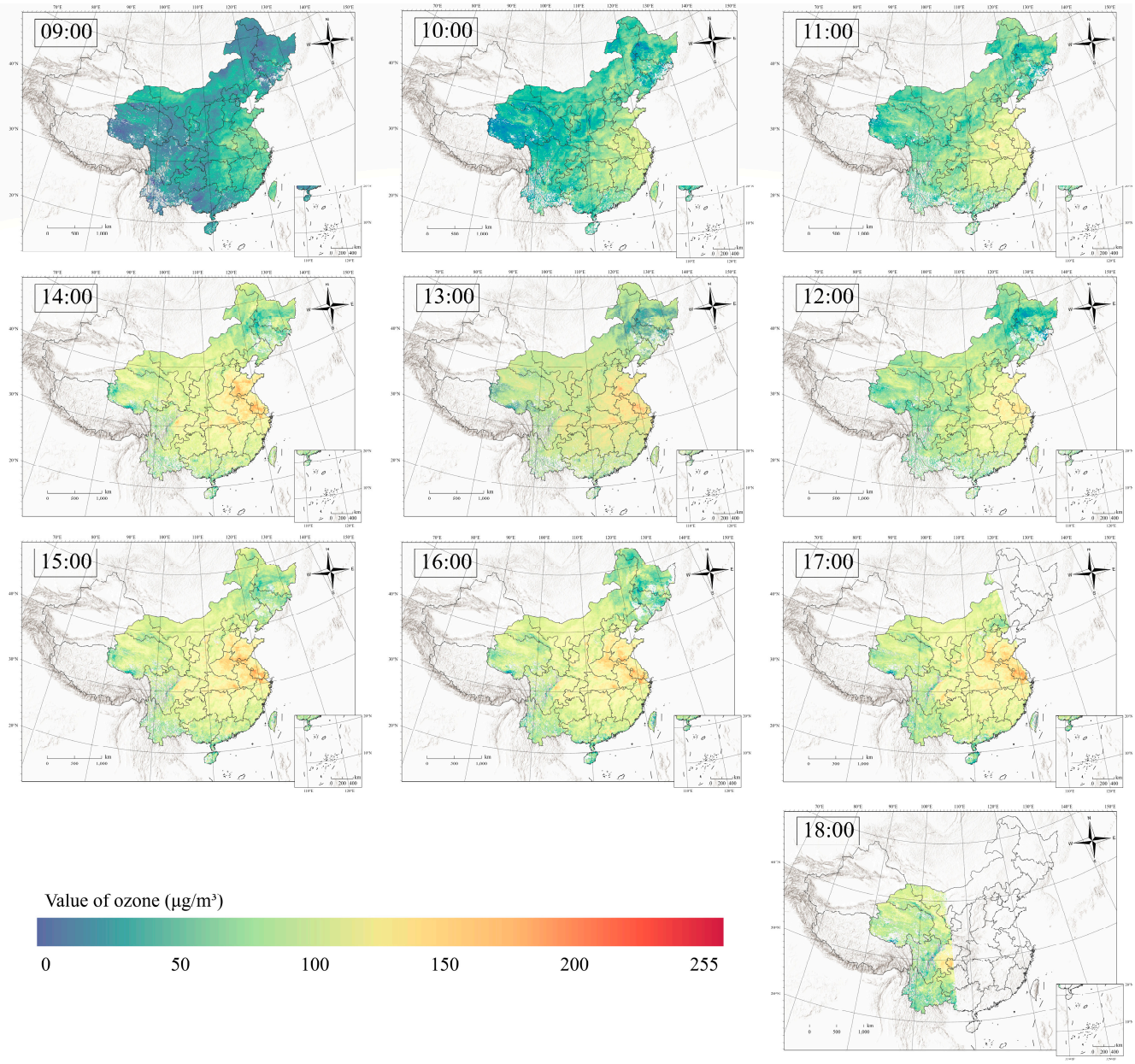

Figure 6. Average multi-day (20 September 2021–26 September 2021) estimations of ground-level O_3 concentrations for each hour at a spatial resolution of 2 km.

Regarding the spatial distribution, we identify high-value areas of O_3 concentrations in eastern coastal regions such as Shandong, Jiangsu, and Zhejiang provinces, where the values concentrated around 210 μg/m^3. High-density regions characterized by substantial industrial, transportation, and residential emissions are responsible for elevated levels of O_3 precursors, including NOx and VOCs [44,45]. Moreover, the culmination of summer characterized by elevated temperatures and increased thunderstorms intensified the increase in ground-level O_3 levels. Conversely, areas at higher latitudes, such as northeast China (Heilongjiang, Jilin, and Liaoning provinces) and the Inner Mongolia Autonomous Region, experienced a slower increase in ground-level O_3 concentrations because of diminished solar radiation caused by higher latitudes. Unfortunately, starting at 16:00, the sun begins to set in Heilongjiang Province. Himawari-8 cannot provide nighttime cloud property data, resulting in partial data gaps after 16:00. This accounts for the significant fluctuations in the standard deviation of ground-level O_3 concentrations between 16:00 and 18:00.

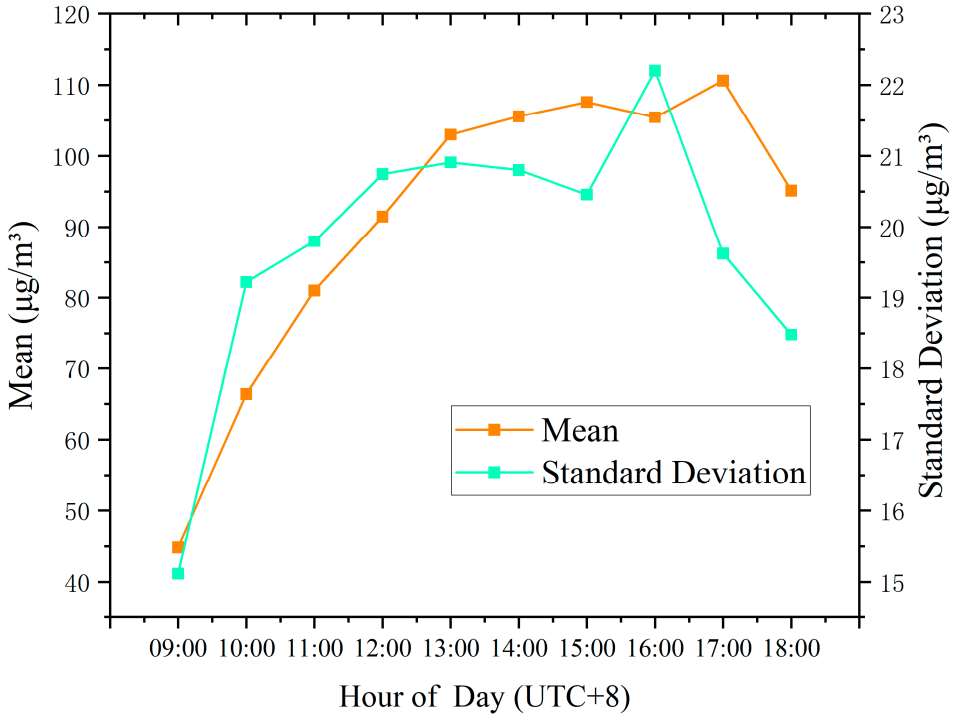

Figure 7. Mean and SD of average multi-day estimation for each hour of ground-level O_3 concentrations.

When considering the temporal aspect, combining these two figures, we observed that the predicted ground-level O_3 concentrations as well as the standard deviation experienced a significant and quick increase from 09:00 to 13:00. This phenomenon was caused by the increasing solar radiation and temperature [46], which facilitate the chemical production of ground-level O_3. Subsequently, the mean ground-level O_3 concentrations reached a steady state of approximately 105 μg/m^3 until sunset, as ground-level O_3 does not disperse quickly before dusk. Regarding the sudden changes at 16:00, the easternmost section of the study area began to be influenced by the sunset, leading to a rapid dissipation of ground-level O_3. At this time, the sunset had a lesser effect on the remaining research area and did not yet cause the dissipation of ground-level O_3. Consequently, the mean of ground-level O_3 concentrations increased. Between 17:00 and 18:00, the influence of the sunset progressively extended to the middle and western areas of the study area. However, due to the limitations of the Himawari satellite, we were unable to collect data after sunset. Consequently, the statistics for both periods do not include the lowest ground-level O_3 concentrations in the eastern portion of the area. As a result, the mean of ground-level O_3 rose in comparison to the prior period.

Finally, we compared our average multi-day (20 September 2021–26 September 2021) estimations with results of the ECMWF's CAMS, as shown in Figure 8. The comparative results indicate that the ozone concentration trends predicted by both models were generally consistent, especially at 14:00 (UTC + 8). Moreover, our high-precision results provide a more detailed reflection of the changes in near-surface ozone concentration. We believe that our study can contribute to the scientific prevention and control of ozone pollution.

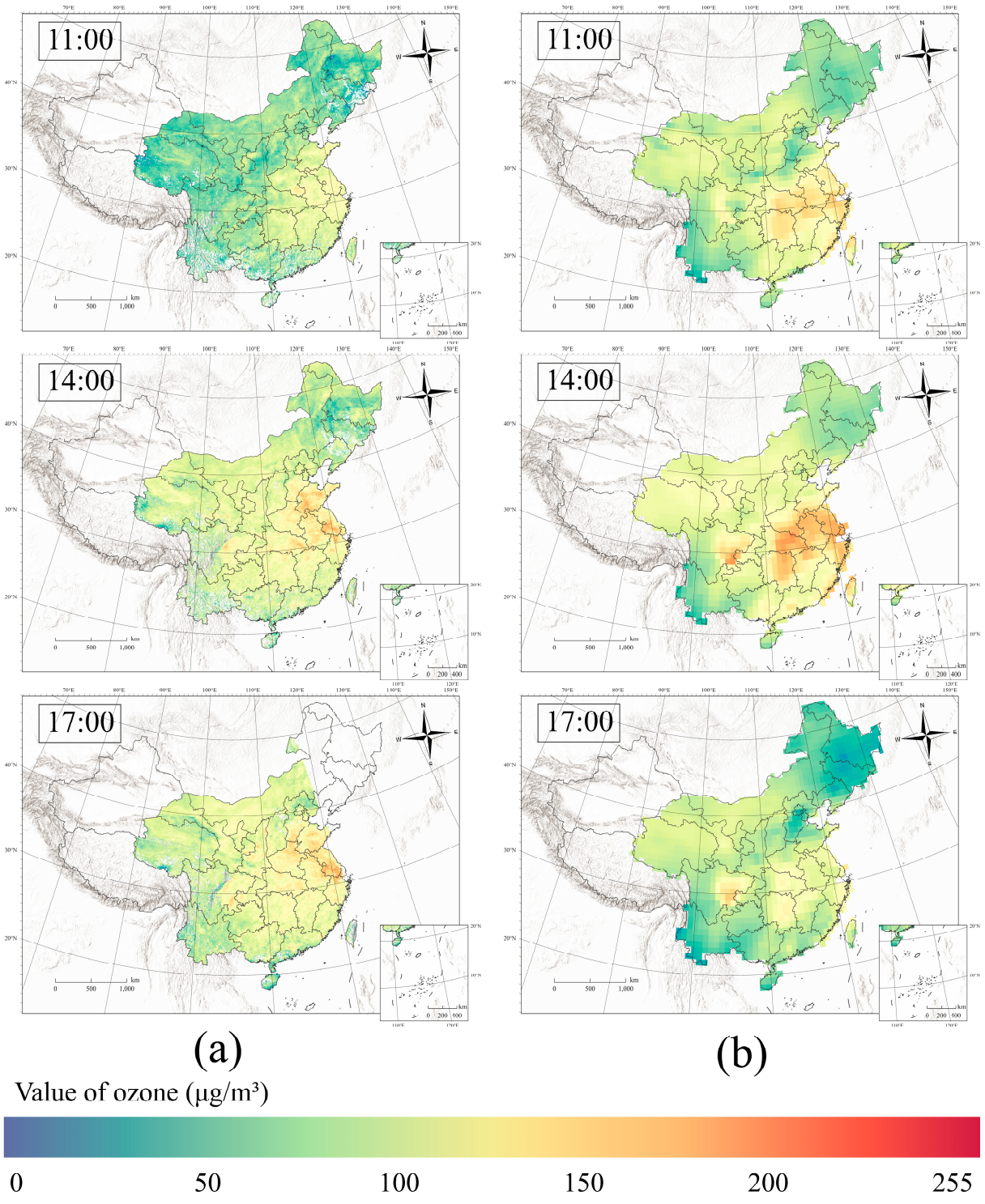

Figure 8. The comparison of estimations between CatBoost model and CAMS. (**a**) The results of the CatBoost model and (**b**) the results of CAMS. Since CAMS estimations are available three-hourly starting from 8:00 (UTC + 8), the comparison involves only three time points.

4. Conclusions

Upon training the machine learning models with the most effective feature combinations and assessing the performance of the four models, we determined that the CatBoost model exhibited optimal performance in this research endeavor. The chosen features exerted a favorable influence on the model's predictions, specifically the features

of T2M, BLH, and data from band 12 of AHI. Afterwards, we employed the CatBoost model to estimate the average multi-day ground-level O_3 concentrations for each hour in the study area. The findings demonstrated a robust association between ground-level O_3 levels and the intensity of solar radiation, with peak values even reaching as high as 210 $\mu g/m^3$, hence presenting a substantial health hazard to inhabitants. Moreover, the spatial distribution of ground-level O_3 concentrations was notably impacted by the extent of human activity. Areas characterized by more concentrated human activity and greater industrial emissions displayed elevated levels of near-ground O_3. We hope that the high-spatiotemporal-resolution estimation in our study will contribute to the scientific management of ground-level O_3.

Author Contributions: Conceptualization, J.C.; investigation, H.D., B.Q. and L.L.; methodology, J.C., H.D. and Z.Z.; writing—original draft, J.C., H.D. and L.L.; writing—review and editing, J.C. and B.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Open Funding of Zhejiang Key Laboratory of Ecological and Environmental Big Data under grant EED-2022-07 and National Natural Science Foundation of China under grant 52079101.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly freely available through the internet. Station data: http://www.cnemc.cn/sssj/ (accessed on 8 December 2023); Himawari data: JAXA Himawari Monitor (P-Tree System) (accessed on 8 December 2023); ERA5-Land data: https://www.ecmwf.int/en/era5-land (accessed on 8 December 2023).

Conflicts of Interest: The authors declare no conflicts of interest. Heng Dong is employee of Zhejiang Spatiotemporal Sophon Bigdata Co., Ltd. The paper reflects the views of the scientists, and not the company.

References

1. Kokhanovsky, A.; Iodice, F.; Lelli l Zschaege, A.; Quattro, N.; Gasbarra, D.; Retscher, C. Retrieval of Total Ozone Column Using High Spatial Resolution Top-of-Atmosphere Measurements by OLCI/S-3 in the Ozone Chappuis Absorption Band over Bright Underlying Surfaces. *J. Quant. Spectrosc. Radiat. Transf.* **2021**, *276*, 107903. [CrossRef]
2. Verstraeten, W.W.; Neu, J.L.; Williams, J.E.; Bowman, K.W.; Worden, J.R.; Boersma, K.F. Rapid Increases in Tropospheric Ozone Production and Export from China. *Nat. Geosci.* **2015**, *8*, 690–695. [CrossRef]
3. Fishman, J.; Ramanathan, V.; Crutzen, P.J.; Liu, S.C. Tropospheric Ozone and Climate. *Nature* **1979**, *282*, 818–820. [CrossRef]
4. Guo, Y.; Zeng, H.; Zheng, R.; Li, S.; Barnett, A.G.; Zhang, S.; Zou, X.; Huxley, R.; Chen, W.; Williams, G. The Association between Lung Cancer Incidence and Ambient Air Pollution in China: A Spatiotemporal Analysis. *Environ. Res.* **2016**, *144*, 60–65. [CrossRef] [PubMed]
5. Zhang, Y.; Ke, L.; Ma, X.; Di, Q. Impact of Ground-Level Ozone Exposure on Sleep Quality and Electroencephalogram Patterns at Different Time Scales. *Environ. Res.* **2023**, *218*, 115025. [CrossRef]
6. Liu, H.; Liu, S.; Xue, B.; Lv, Z.; Meng, Z.; Yang, X.; Xue, T.; Yu, Q.; He, K. Ground-Level Ozone Pollution and Its Health Impacts in China. *Atmos. Environ.* **2018**, *173*, 223–230. [CrossRef]
7. WHO. *WHO Global Air Quality Guidelines: Particulate Matter (PM2. 5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*; World Health Organization: Geneva, Switzerland, 2021.
8. The Ministry of Ecology and Environment of the People's Republic of China. *Report on China's Ecological and Environmental Status in 2021*; The Ministry of Ecology and Environment of the People's Republic of China: Beijing, China, 2021.
9. Chen, J.; Shen, H.; Li, X.; Li, T.; Wei, Y. Ground-Level Ozone Estimation Based on Geo-Intelligent Machine Learning by Fusing in-Situ Observations, Remote Sensing Data, and Model Simulation Data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102955. [CrossRef]
10. Wang, W.; Liu, X.; Bi, J.; Liu, Y. A Machine Learning Model to Estimate Ground-Level Ozone Concentrations in California Using TROPOMI Data and High-Resolution Meteorology. *Environ. Int.* **2022**, *158*, 106917. [CrossRef]
11. Wang, H.-W.; Li, X.-B.; Wang, D.; Zhao, J.; He, H.; Peng, Z.-R. Regional Prediction of Ground-Level Ozone Using a Hybrid Sequence-to-Sequence Deep Learning Approach. *J. Clean. Prod.* **2020**, *253*, 119841. [CrossRef]
12. Kim, S.-W.; Yoon, S.-C.; Won, J.-G.; Choi, S.-C. Ground-Based Remote Sensing Measurements of Aerosol and Ozone in an Urban Area: A Case Study of Mixing Height Evolution and Its Effect on Ground-Level Ozone Concentrations. *Atmos. Environ.* **2007**, *41*, 7069–7081. [CrossRef]

13. He, S.; Dong, H.; Zhang, Z.; Yuan, Y. An Ensemble Model-Based Estimation of Nitrogen Dioxide in a Southeastern Coastal Region of China. *Remote Sens.* **2022**, *14*, 2807. [CrossRef]
14. He, Q.; Ye, T.; Chen, X.; Dong, H.; Wang, W.; Liang, Y.; Li, Y. Full-Coverage Mapping High-Resolution Atmospheric CO2 Concentrations in China from 2015 to 2020: Spatiotemporal Variations and Coupled Trends with Particulate Pollution. *J. Clean. Prod.* **2023**, *428*, 139290. [CrossRef]
15. Bessho, K.; Date, K.; Hayashi, M.; Ikeda, A.; Imai, T.; Inoue, H.; Kumagai, Y.; Miyakawa, T.; Murata, H.; Ohno, T.; et al. An Introduction to Himawari-8/9—Japan's New-Generation Geostationary Meteorological Satellites. *J. Meteorol. Soc. Japan Ser. II* **2016**, *94*, 151–183. [CrossRef]
16. Boynard, A.; Clerbaux, C.; Coheur, P.-F.; Hurtmans, D.; Turquety, S.; George, M.; Hadji-Lazaro, J.; Keim, C.; Meyer-Arnek, J. Measurements of Total and Tropospheric Ozone from IASI: Comparison with Correlative Satellite, Ground-Based and Ozonesonde Observations. *Atmos. Chem. Phys.* **2009**, *9*, 6255–6271. [CrossRef]
17. Clerbaux, C.; Boynard, A.; Clarisse, L.; George, M.; Hadji-Lazaro, J.; Herbin, H.; Hurtmans, D.; Pommier, M.; Razavi, A.; Turquety, S.; et al. Monitoring of Atmospheric Composition Using the Thermal Infrared IASI/MetOp Sounder. *Atmos. Chem. Phys.* **2009**, *9*, 6041–6054. [CrossRef]
18. Peng, Z.; Letu, H.; Wang, T.; Shi, C.; Zhao, C.; Tana, G.; Zhao, N.; Dai, T.; Tang, R.; Shang, H.; et al. Estimation of Shortwave Solar Radiation Using the Artificial Neural Network from Himawari-8 Satellite Imagery over China. *J. Quant. Spectrosc. Radiat. Transf.* **2020**, *240*, 106672. [CrossRef]
19. Chen, B.; Wang, Y.; Huang, J.; Zhao, L.; Chen, R.; Song, Z.; Hu, J. Estimation of Near-Surface Ozone Concentration and Analysis of Main Weather Situation in China Based on Machine Learning Model and Himawari-8 TOAR Data. *Sci. Total Environ.* **2023**, *864*, 160928. [CrossRef]
20. Wang, Y.; Yuan, Q.; Zhu, L.; Zhang, L. Spatiotemporal Estimation of Hourly 2-km Ground-Level Ozone over China Based on Himawari-8 Using a Self-Adaptive Geospatially Local Model. *Geosci. Front.* **2022**, *13*, 101286. [CrossRef]
21. Shaohu, Z.; Xiaoyu, Y.; Zhengqiang, L.; Zhongting, W.; Yuhuan, Z.; Yu, W.; Chunyan, Z.; Pengfei, M. Advances of Ozone Satellite Remote Sensing in 60 Years. *J. Remote Sens.* **2022**, *26*, 817–833.
22. Zhen, L.; Heng, D.; Zili, D.; Lan, L.; Sicong, H. Estimation of Near-Ground Ozone with High Spatio-Temporal Resolution in the Yangtze River Delta Region of China Based on a Temporally Ensemble Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7051–7061. [CrossRef]
23. Henze, D.K.; Hakami, A.; Seinfeld, J.H. Development of the Adjoint of GEOS-Chem. *Atmos. Chem. Phys.* **2007**, *7*, 2413–2433. [CrossRef]
24. Zhang, L.; Jacob, D.J.; Downey, N.V.; Wood, D.A.; Blewitt, D.; Carouge, C.C.; van Donkelaar, A.; Jones, D.B.; Murray, L.T.; Wang, Y. Improved Estimate of the Policy-Relevant Background Ozone in the United States Using the GEOS-Chem Global Model with 1/2 × 2/3 Horizontal Resolution over North America. *Atmos. Environ.* **2011**, *45*, 6769–6776. [CrossRef]
25. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.-M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS Reanalysis of Atmospheric Composition. *Atmos. Chem. Phys.* **2019**, *19*, 3515–3556. [CrossRef]
26. Amnuaylojaroen, T.; Barth, M.; Emmons, L.; Carmichael, G.; Kreasuwun, J.; Prasitwattanaseree, S.; Chantara, S. Effect of Different Emission Inventories on Modeled Ozone and Carbon Monoxide in Southeast Asia. *Atmos. Chem. Phys.* **2014**, *14*, 12983–13012. [CrossRef]
27. Lee, G.; Lee, J.-H. Spatial Correlation Analysis Using the Indicators of the Anthropocene Focusing on Atmospheric Pollution: A Case Study of Seoul. *Ecol. Indic.* **2021**, *125*, 107535. [CrossRef]
28. Duan, X.; Cao, N.; Wang, X.; Zhang, Y.; Liang, J.; Yang, S.; Song, X. Analysis of Near-Surface Ozone Concentrations in China in 2015. *Environ. Sci.* **2017**, *38*, 4976–4982.
29. Gogeri, I.; Gouda, K.C.; Aruna, S.T. Spatio-Temporal Analysis of Air Pollution Dynamics over Bangalore City during Second Wave of COVID-19. *Nat. Hazards Res.* **2023**. [CrossRef]
30. Kerckhoffs, J.; Wang, M.; Meliefste, K.; Malmqvist, E.; Fischer, P.; Janssen, N.A.; Beelen, R.; Hoek, G. A National Fine Spatial Scale Land-Use Regression Model for Ozone. *Environ. Res.* **2015**, *140*, 440–448. [CrossRef]
31. Felder, M.; Sehnke, F.; Kaifel, A. Combined Ozone Retrieval from METOP Sensors Using META-Training of Deep Neural Networks. *Proc. ESA Living Planet Symp.* **2013**, *722*, 219.
32. Zhan, Y.; Luo, Y.; Deng, X.; Grieneisen, M.L.; Zhang, M.; Di, B. Spatiotemporal Prediction of Daily Ambient Ozone Levels across China Using Random Forest for Human Exposure Assessment. *Environ. Pollut.* **2018**, *233*, 464–473. [CrossRef]
33. Li, R.; Cui, L.; Hongbo, F.; Li, J.; Zhao, Y.; Chen, J. Satellite-Based Estimation of Full-Coverage Ozone (O_3) Concentration and Health Effect Assessment across Hainan Island. *J. Clean. Prod.* **2020**, *244*, 118773. [CrossRef]
34. Li, Y.; Qin, K.; Li, D.; Pan, W.; He, Q. Estimation of Ground-Level Ozone Concentration Based on Gradient Boosting Regression Trees Algorithm. *China Environ. Sci.* **2020**, *40*, 997–1007.
35. Li, Y. Remote Sensing Estimation of Ground-Level O_3 Concentrations in China Based on Gradient Boosting Regression Trees. Master's Thesis, China University of Mining and Technology, Beijing, China, 2020.
36. Zhao, N.; Lu, Y. Remote Estimation of Ground-Level Ozone Concentration Based on the XGBoost Algorithm. *Acta Sci. Circumstantiae* **2022**, *42*, 95–108.
37. Jingping, W.; Xiaodan, W.; Dujuan, M.; Jianguang, W.; Qing, X. Machine Learning-Based Remote Sensing Inversion: Analysis of Uncertainty Factors. *J. Remote Sens.* **2023**, *27*, 790–801.

38. Zhou, X.; Zhu, X.; Dong, Z.; Guo, W. Estimation of Biomass in Wheat Using Random Forest Regression Algorithm and Remote Sensing Data. *Crop J.* **2016**, *4*, 212–219.
39. Yan, X.; Gou, X.; Yang, J.; Zhao, W.; Xu, Q.; Liu, Y. The Variety of Ozone and Its Relationship with Meteorological Conditions in Typical Cities in China. *Plateau Meteorol.* **2020**, *39*, 416–430.
40. Liu, J.; He, C.; Zhao, S.; Zhu, J.; Wang, W.; Wang, L.; Wang, Y. Variations in Ozone Concentration in Seven Regions under Different Temperature and Humidity Conditions. *Huanjing Kexue* **2023**, *44*, 5392–5399.
41. Chen, Z.; Li, R.; Chen, D.; Zhuang, Y.; Gao, B.; Yang, L.; Li, M. Understanding the Causal Influence of Major Meteorological Factors on Ground Ozone Concentrations across China. *J. Clean. Prod.* **2020**, *242*, 118498. [CrossRef]
42. Wang, T.; Xue, L.; Brimblecombe, P.; Lam, Y.F.; Li, L.; Zhang, L. Ozone Pollution in China: A Review of Concentrations, Meteorological Influences, Chemical Precursors, and Effects. *Sci. Total Environ.* **2017**, *575*, 1582–1596. [CrossRef]
43. Tang, G.; Liu, Y.; Huang, X.; Wang, Y.; Hu, B.; Zhang, Y.; Song, T.; Li, X.; Wu, S.; Li, Q.; et al. Aggravated Ozone Pollution in the Strong Free Convection Boundary Layer. *Sci. Total Environ.* **2021**, *788*, 147740. [CrossRef]
44. Zang, X.; Lu, Y.; Yao, H.; Li, F.; Zhang, S. Study of the Spatiotemporal Distribution Characteristics of Major Atmospheric Pollutants in China. *Ecol. Environ. Sci.* **2015**, *24*, 1322.
45. He, S.; Yuan, Y.; Wang, Z.; Luo, L.; Zhang, Z.; Dong, H.; Zhang, C. Machine Learning Model-Based Estimation of XCO2 with High Spatiotemporal Resolution in China. *Atmosphere* **2023**, *14*, 436. [CrossRef]
46. Kou, W.; Gao, Y.; Zhang, S.; Cai, W.; Geng, G.; Davis, S.; Wang, H.; Guo, X.; Cheng, W.; Zeng, X.; et al. High downward surface solar radiation conducive to ozone pollution more frequent under global warming. *Sci. Bull* **2023**, *68*, 388–392. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.