remote sensing

MDPI

Article

When a Generalized Linear Model Meets Bayesian Maximum Entropy: A Novel Spatiotemporal Ground-Level Ozone Concentration Retrieval Method

Yingying Mei [1], Jiayi Li [2,*], Deping Xiang [3] and Jingxiong Zhang [4]

[1] School of Sociology, Wuhan University, Wuhan 430072, China; myy2014@whu.edu.cn
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[3] School of Sociology, Huazhong University of Science and Technology, Wuhan 430074, China; 2019010080@hust.edu.cn
[4] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; jxzhang@whu.edu.cn
* Correspondence: zjjercia@whu.edu.cn

Abstract: In China, ground-level ozone has shown an increasing trend and has become a serious ambient pollutant. An accurate spatiotemporal distribution of ground-level ozone concentrations (GOCs) is urgently needed. Generalized linear models (GLMs) and Bayesian maximum entropy (BME) models are practical for predicting GOCs. However, GLMs have limited capacity to capture temporal variations and can miss some short-term and regional patterns, while the performance of BME models may degrade in cases of sparse or imperfect monitoring networks. Thus, to predict nationwide 1 km monthly average GOCs for China, we designed a novel hybrid model containing three modules. (1) A GLM was established to accurately describe the variability in GOCs in the space domain. (2) A BME model incorporating GLM residuals was employed to capture the temporal variability of GOCs in detail. (3) A combination of GLM and BME models was developed based on the specific broad range of each submodel. According to the cross-validation results, the hybrid model exhibited superior performance, with coefficient of determination (R^2) values of 0.67. The predictive performance of the large-scale and high-resolution hybrid model is superior to that in previous studies. The nationwide spatiotemporal variability of the GOCs derived from the hybrid model shows that they are valuable indicators for ground-level ozone pollution control and prevention in China.

Keywords: ground-level ozone; national scale; China; spatiotemporal distribution; hybrid model

Citation: Mei, Y.; Li, J.; Xiang, D.; Zhang, J. When a Generalized Linear Model Meets Bayesian Maximum Entropy: A Novel Spatiotemporal Ground-Level Ozone Concentration Retrieval Method. *Remote Sens.* **2021**, *13*, 4324. https://doi.org/10.3390/rs13214324

Academic Editors: Simone Lolli and Daniele Bortoli

Received: 2 September 2021
Accepted: 20 October 2021
Published: 27 October 2021

1. Introduction

In recent years, increasing ground-level ozone concentrations (GOCs) have attracted worldwide attention because of their adverse effects on human health, climate, and vegetation [1,2]. To better understand these adverse effects, an accurate and high-resolution GOC distribution is urgently needed.

The prediction of GOCs has become a research topic of interest in the atmospheric environment community. In view of model construction (Table 1), two broad types of models have been employed to predict GOCs: deterministic models and statistical models. Deterministic models, such as air quality models [2,3], weather research and forecasting models [4,5], and chemical transport models [6], can predict GOCs based on the theoretical description of ozone formation processes, but the computation process is relatively complex [7], and it is difficult to obtain high-resolution products (e.g., 1 km).

Table 1. Model survey of recent studies on GOC prediction.

Model		Study Area	Temporal/ Spatial Resolution	R^2	*RMSE* (μg/m^3)	Reference
Deterministic model		Site-specific	Daily/-	-	-	[8]
		National	Annual/-	-	-	[9]
		City	Daily/750 m	-	-	[10]
		City	Daily/50 km	-	-	[11]
Statistical model	GLM	Site-specific	-	0.56–0.80	-	[12]
		Site-specific	-	0.35–0.81	-	[13]
		Site-specific	-	0.34–0.71	8.07–14.24	[14]
	BME	City	Monthly/1 km	0.65	7.06	[15]
		City	Monthly/-	-	-	[16]
		City	Daily/-	-	-	[6]
	Others	National	Daily/-	0.74	7.2	[17]
		Site-specific	Seasonal/-	-	-	[18]
		National	Monthly/0.1°	0.60–0.87	-	[19]
		Site-specific	Daily/-	-	18.4–42.7	[20]

The explanatory variables used to predict GOC include meteorological conditions and environmental pollutants. First, as the most frequently used variables, meteorological conditions (e.g., temperature, wind and precipitation) strongly influence ozone formation and deposition. Specifically, temperature, a proxy for solar radiation, accelerates/decelerates the speed of photochemical reactions in GOCs [2,21]; wind usually affects GOCs by influencing atmospheric mixing, dispersion, and transport [15]; and precipitation leads to a reduction in photochemical reaction efficiency [22]. Second, several environmental pollutants originating from human activity [13] and natural sources [23] are also applied to predict GOCs. According to the source appointment performed by the Ministry of Environment, ground-level ozone is a by-product of many human activities, including traffic emissions, coal combustion, and industrial emissions [24]. Pollutants emitted by natural sources are considered in prediction models due to their correlation with photochemical reactions in ozone formation [20]. In general, an increase in explanatory variables can improve the predictive performance of a model. However, explanatory variables across a broad geographic area on a large scale are not easy to acquire at present [25].

In China, urbanization and industrialization have led to severe ground-level ozone pollution [26,27]. This emerging severity of ozone pollution presents a new challenge for emission control strategies. According to a survey of mainstream research, most studies have focused on site-specific predictions [9,13,14,25,28] or city scales [5,10,11,22,29]. Meanwhile, in situ measurements of GOCs in China remain insufficient to assess the nationwide effects of ozone pollution [9]. Thus, an increasing number of studies have focused on the prediction of GOCs in some areas of China (such as cities or provinces) [25,29]. A few scholars have tried to build models at the national scale for China, such as Zhan et al. (2018) [30] and Liu et al. (2020) [19]. However, such studies are mostly performed at a relatively rough resolution (i.e., 0.1 degrees or even coarser).

In recent years, statistical models, such as generalized linear models (GLMs) [12,14] and Bayesian maximum entropy (BME) models [8,31], have become more practical in GOC prediction by employing correlations between ozone measurements and related explanatory variables [12,32]. GLM can accurately describe the variability in GOCs in the space domain by using a nonlinear regression framework [12] but has limited capacity to capture temporal variations and can miss some short-term and regional patterns [14]. BME is an interpolation method that assigns a series of weights to observed monitoring site data to compute concentrations at measurement sites. While a BME model is capable of describing the temporal variability of GOCs in detail, its performance degrades in cases of sparse or imperfect monitoring networks [15]. Other statistical models, such as machine learning techniques and land use regression models, were also used to predict air pollution

concentrations. However, "black-box like" machine learning methods are insufficient to explain and analyze the relationship between input and output variables [17,19,20], and land use regression models are usually conducted on a relatively small geographical scale [15].

According to recent studies, combining a coarse estimate from a regression model and a refined variation from an interpolation model may have merit [15,33]. For example, the combination of a land use regression (LUR) model and a BME model could effectively increase the predictive accuracy for pollutant concentrations, such as GOCs in Quebec [15] and $PM_{2.5}$ concentrations in China [33]. However, LUR models are somewhat inadequate in capturing the nonlinear relationship between GOCs and related explanatory variables [34]. Inspired by these works and the advantages of GLMs and BME models, we designed a novel hybrid model that combines a GLM and a BME model to achieve a satisfactory estimation of monthly average concentrations at a high spatial resolution (i.e., 1 km) across China in 2018. According to the predictive abilities and data accessibility, land surface temperature (LST), bias-corrected total precipitation (BCTP), total column production of precipitation (TCPP), 2 m specific humidity (SH), relative humidity after moisture (RHAM), road density (RD), longitude (LON), latitude (LAT), and day number sequence (DNS) were utilized. Cross-validation was applied to test the model performance. Using the hybrid model, we derived the nationwide spatiotemporal distributions of monthly GOC and conducted spatiotemporal pattern analysis, which are helpful to control, understand, and prevent ground-level ozone pollution in China.

2. Study Areas

This study takes mainland China as the research area. Large-scale real-time GOCs have been regularly monitored since 2013 via a national air quality monitoring network in China [9,30]. The coordinates of the national monitoring sites and the hourly ground-level ozone records from 13 May 2014, to 1 August 2019, are acquired from the China National Environmental Monitoring Center (CNEMC). The monitoring sites are mainly located in East, Central, and South China, but they are relatively sparse in North, Northeast, Northwest, and Southwest China [35] (Figure 1). We estimated the daily average concentration of ozone for each site when at least 75% of the hourly measurements were available, according to [15].

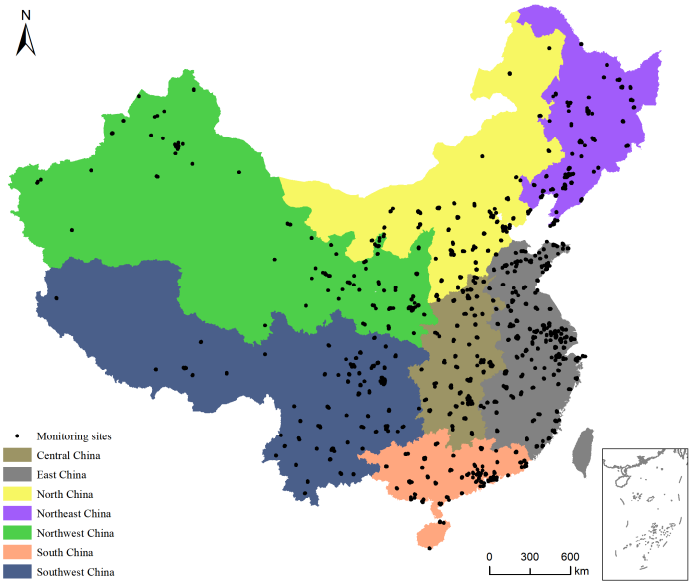

Figure 1. Geographic locations of the monitoring sites in China.

3. Hybrid Model: GLM + BME

In this study, a hybrid model was developed to predict nationwide monthly average GOCs at a 1 km spatial resolution. The three major modules (see Figure 2) constituting the proposed hybrid model are described in Sections 3.2–3.4. Section 3.2 describes the variability in GOCs in the space domain by using a GLM. Then, a BME model incorporating information from the GLM was employed to capture the temporal variations in GOCs in Section 3.3. Finally, the information from GLM and BME models were combined to obtain more accurate predictions based on the specific broad range of each submodel in Section 3.4.

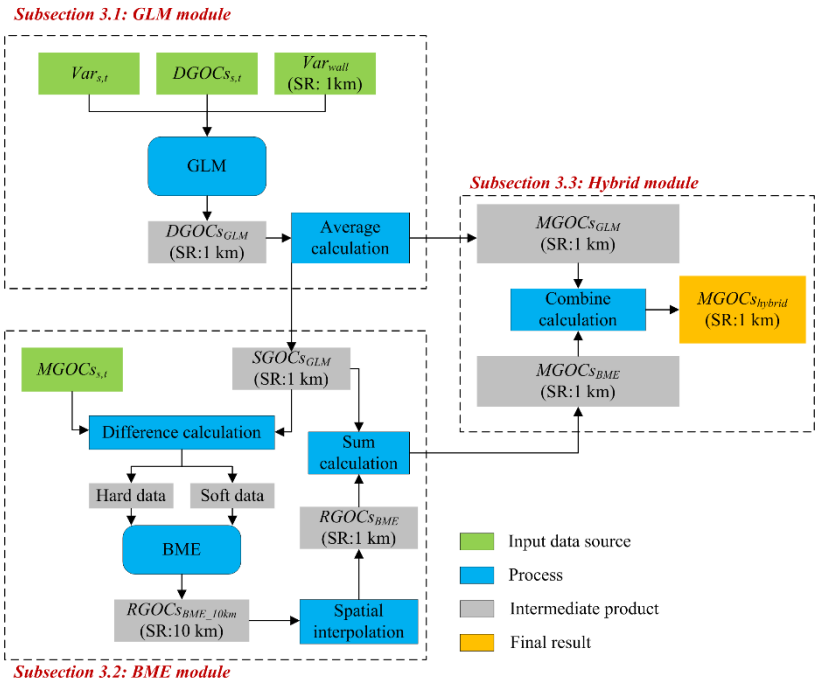

Figure 2. Flow chart of the hybrid model (SR: spatial resolution).

3.1. Materials

The list of utilized explanatory variables was summarized (see Table 2). Multisource geospatial data, including LST, BCTP, TCPP, SH, RHAM, and RD, were obtained and resampled to a uniform grid cell (i.e., 1 km × 1 km) by using ArcGIS 10.5, ENVI 4.7, and MATLAB R2014a. For example, LST (spatial resolution: 1 km × 1 km) was derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) product MYD11A1 [36]. BCTP, TCPP, SH, and RHAM (spatial resolution of both: 0.667° × 0.5°) were obtained from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) and resampled by nearest-neighbor interpolation [37]. All the meteorological products ranged from 13 May 2014, to 1 August 2019. The RD value D was calculated based on Equation (1).

$$D = L/A \quad (1)$$

where L is the total length of the major roads extracted from OpenStreetMap (OSM) (i.e., motorway, trunk, primary and secondary highways) within a circular area with a 1 km radius, and A is the circular area. The multisource geospatial data, together with geolocation information (i.e., LON and LAT) and temporal information (i.e., DNS), were used as explanatory variables in building the hybrid model. All the data (including ground-level ozone measurements and explanatory variables) were matched by their grid cell ID and DNS for model development and validation.

Table 2. List of explanatory variables.

Explanatory Variable	Unit	Spatial Resolution	Temporal Resolution	Preprocessing Method
LST	°C	1 km × 1 km	Day	Spatial overlay
BCTP	$kgm^{-2}s^{-1}$	0.667° × 0.5°	Day	Nearest-neighbor interpolation
TCPP	$kgm^{-2}s^{-1}$	0.667° × 0.5°	Day	Nearest-neighbor interpolation
SH	$kgkg^{-1}$	0.667° × 0.5°	Day	Nearest-neighbor interpolation
RHAM	1	0.667° × 0.5°	Day	Nearest-neighbor interpolation
RD	km/km^2	Polyline	Year	$D = L/A$
LON	°	NA	NA	NA
LAT	°	NA	NA	NA
DNS	NA	NA	NA	NA

LST: land surface temperature, BCTP: bias-corrected total precipitation, TCPP: total column production of precipitation, SH: 2 m specific humidity, RHAM: relative humidity after moisture, RD: road density, LON: longitude, LAT: latitude and DNS: day number sequence.

3.2. GLM Module

The GLM module is performed to produce monthly and seasonal average GOCs at a 1 km spatial resolution (i.e., $MGOCs_{GLM}$ and $SGOCs_{GLM}$). Daily average concentrations of eligible site-days ($DGOCs_{s,t}$), station-days explanatory variables ($Var_{s,t}$), and wall-to-wall explanatory variables at a 1 km spatial resolution (Var_{wall}) were used as data sources. This step mainly included the building of GLM, variable selection, and prediction of 1 km daily nationwide GOCs ($DGOCs_{GLM}$).

First, a GLM for predicting daily GOCs was constructed based on the nonlinear relationship between daily concentrations and explanatory variables of eligible site days [12]:

$$g(\mu_{d_m}) = \log \frac{\mu_{d_m}}{1 - \mu_{d_m}} = \mathbf{F}^T \boldsymbol{\beta} \quad (2)$$

where μ_{d_m} is the daily average concentration of monitoring site m on day d, $g(\cdot)$ is the log link function, and $\boldsymbol{\beta}$ represents the regression coefficient vector. $\mathbf{F} = (F_0 = 1, F_1, \ldots, F_n)^T$ is a vector that represents the explanatory variables in the regression model, n is the number of explanatory variables except for the constant term $F_0 = 1$, and T denotes transpose. The GLM analysis was carried out using the glm function in Rstudio 3.5.1.

Second, a variable selection procedure based on the Akaike information criterion (AIC) was utilized to identify the variables included in the final GLM. The variables were added one by one in the initial model. At each step, the significance of the added variable to the model was tested. Any variable whose p-value was higher than 0.05 was excluded. According to the selection procedure, LST, BCTP, RHAM, RD, LON, LAT, and DNS were left in the final GLM. Variable selection was conducted using Rstudio 3.5.1.

Third, using the wall-to-wall explanatory variables (Var_{wall}), spatiotemporal characteristics of daily GOCs ($DGOCs_{GLM}$) were predicted by the final GLM. Then, $DGOCs_{GLM}$ were further aggregated into $MGOCs_{GLM}$ and $SGOCs_{GLM}$. One of the $MGOCs_{GLM}$ is shown in Figure 3. As shown in Figure 3, the monthly average GOCs in November 2018 exhibited spatial heterogeneity across China and high monthly average concentrations tended to occur in South, Central, and East China.

3.3. BME Module

The spatiotemporal distribution of GOCs derived from the GLM model (Section 3.2) described spatial variability characteristics in GOCs well; however, even when a time variable (i.e., DNS) was included, they were deficient in capturing temporal variability in detail [12]. Thus, a BME model incorporating information from the GLM was employed to obtain the 1 km monthly average concentrations ($MGOCs_{BME}$). The data sources included the measured monthly average concentrations of monitoring sites ($MGOCs_{s,t}$) and the $DGOCs_{GLM}$ derived from Section 3.2. This step mainly included the establishment of the BME model based on hard data and soft data, the construction of an empirical spatiotemporal covariance model, and the prediction of monthly residual concentrations

at a 1 km spatial resolution ($RGOCs_{BME}$). Hard data are usually represented by accurate measurements obtained from real-time observation devices and numerical simulations. Soft data refer to information that can be used to improve estimates by compensating for the limited amount of measured data, including incomplete and qualitative observations, which are usually expressed in terms of interval values, probability statements, empirical charts, etc.

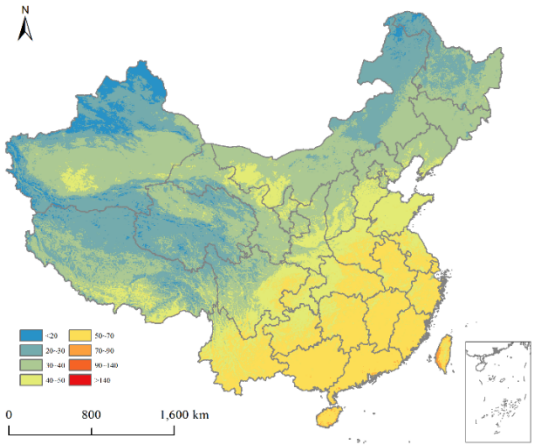

Figure 3. Spatial distributions of the predicted nationwide monthly average GOCs in November 2018 based on the GLM.

First, a BME model was constructed by using MATLAB R2014a and BMElib 2.0b [38].

$$Z_{BME}(R(s,t)) = Z_{BME}(R(p)) = \int pf(p)dp \tag{3}$$

where Z_{BME} represents the GOCs, $R(s,t) = R(p)$ denotes a spatiotemporal random field (RF), p denotes a spatiotemporal point, s is the geographical location, and t is the temporal information. The posterior probability density function $f(p)$ is obtained from the prior probability density function f_G, which can be modeled as follows [39]:

$$f(p) = f_G(p|p_{hard}, p_{soft}) = \frac{f_G(p, p_{hard}, p_{soft})}{f_G(p_{hard}, p_{soft})} \tag{4}$$

where p_{hard} refers to hard data and p_{soft} refers to soft data.

In this paper, the hard data and soft data were obtained from Equations (5) and (6), respectively:

$$p_{hard} = CM(s,t) - SGOCs_{GLM}(s) \tag{5}$$

$$p_{soft} = DM(s,t) - SGOCs_{GLM}(s) \tag{6}$$

where $CM(s,t)$ is the measured monthly average GOC of monitoring site s at time t, $DM(s,t)$ is the distribution of the measured monthly average GOC described by a Gaussian probability density function, and $SGOCs_{GLM}(s)$ is the seasonal average GOC of monitoring site s derived from $DGOCs_{GLM}$.

Second, an empirical spatiotemporal covariance model based on hard data and soft data was used to describe the stochastic processes affecting GOCs. A nested theoretical model consisting of two components (i.e., $c_1(\bullet)$ and $c_2(\bullet)$) was applied to fit the empirical covariance model.

$$c(h,\tau) = C1 \bullet c_1(h,\tau,a_{s1},a_{t1}) + C2 \bullet c_2(h,\tau,a_{s2},a_{t2}) \tag{7}$$

where h is the spatial lag; τ is the temporal lag; $C1$ and $C2$ are the sill coefficients of the two components, respectively; a_{s1} and a_{t1} are the space and time ranges of the first

component, respectively; and a_{s2} and a_{t2} are the space and time ranges of the second component, respectively.

Figure 4 shows that nested theoretical covariance model well represents the empirical spatiotemporal covariance. As shown in Figure 4, the first component (sill = 0.9) described most of the variability; this component consists of an exponential model in space with a range of approximately 1 DD (decimal degrees), approximately 111 km on the Earth's surface, and a spherical model in time with a range of 6 months. The second component (sill = 0.1) consists of a spherical model in space with a range of 5 DD, approximately 555 km on the Earth's surface, and an exponential model in time with a range of 4 months. The short-range interactions addressed by the first component of the fitted theoretical covariance model are consistent with the spatiotemporal scale of emissions from traffic or industry, whereas the long-range interactions described by the second component are consistent with the scale of transport and dispersion over regional domains affected by weather patterns and meteorological conditions [40].

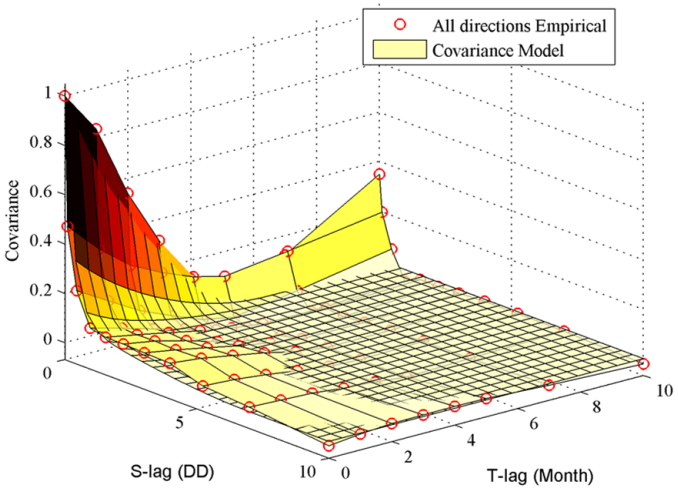

Figure 4. Observed (red circles) and fitted covariance of the residual GOCs (all direction empirical refers to the isotropy assumption, which means the same meaning in all directions).

Third, considering the computational cost over such a large geographical area, we used the BME model to predict the monthly residual concentrations at a 10 km spatial resolution ($RGOCs_{BME\_10km}$), which indicated the difference between measured monthly average GOCs and the seasonal average GOCs derived from the GLM. Then, we resampled each of the $RGOCs_{BME\_10km}$ to 1 km × 1 km grid cells through spatial interpolation to derive the $RGOCs_{BME}$. Finally, $SGOCs_{GLM}$ was added back to $RGOCs_{BME}$ to obtain $MGOCs_{BME}$.

3.4. Hybrid Module

In general, models combining a regression model and interpolation of regression residuals could provide more accurate predictions for pollutant concentrations [33]. However, in areas without extensive monitoring sites, the strong uncertainty existing in interpolation may reduce the prediction accuracy of the BME model [33,41]. Thus, this step was to obtain more accurate predictions for monthly average GOCs at a 1 km spatial resolution ($MGOCs_{hybrid}$). We constructed the hybrid module based on the GLM and the BME model to better predict the spatiotemporal variability of GOCs at a high spatial resolution. The data sources (i.e., $MGOCs_{GLM}$ and $RGOCs_{BME}$) were derived from Sections 3.2 and 3.3, respectively. This step mainly included the elucidation of the specific distance and the combination of information based on the specific broad range of each submodel.

First, the monitoring sites were categorized into seven groups according to the minimum spatial lag between each monitoring site and other monitoring sites. The seven spatial lag classes were defined as 0–20 km, 20–40 km, 40–60 km, 60–80 km, 80–100 km, 100–120 km, and >120 km. $MGOCs_{GLM}$ and $MGOCs_{BME}$ were calculated for each measured monthly

average GOC. For each spatial lag class, we calculated the percent change in R^2 from the GLM to the BME model (e.g., $\{[R^2_{\mathrm{BME}}\text{-}R^2_{\mathrm{GLM}}]/R^2_{\mathrm{GLM}}\} \times 100$) (Figure 5). A positive value indicates that the BME model performs better than the GLM; otherwise, the GLM is superior. As shown in Figure 5, the 100 km corresponding to the inflection point was elucidated as the specific distance.

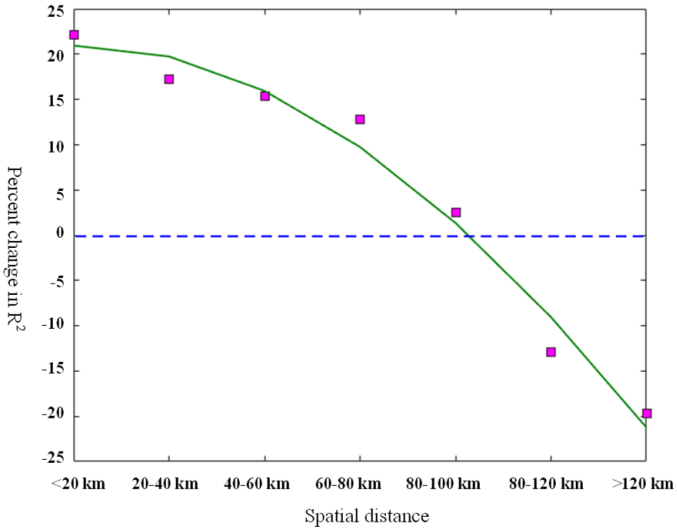

Figure 5. Percent change in R^2 from the GLM to the BME model for different spatial distance.

Second, we combined the information from GLM and BME models according to the distance between the measurement point and its closest monitoring site (*dist*).

$$MGOCs_{hybrid} = \begin{cases} MGOCs_{BME} & dist \le dist_0 \\ MGOCs_{GLM} & dist > dist_0 \end{cases} \quad (8)$$

where $dist_0$ is the specific distance within which BME outperformed the GLM and beyond which the GLM was superior to the BME. According to Equation (8), the hybrid model was constructed by combining the information from the BME model up to approximately 100 km from a monitoring site and information from the GLM beyond that distance.

3.5. Accuracy Evaluation

A site-based 10-fold cross-validation method was utilized to evaluate the predictive performance of different models. The monitoring sites were randomly partitioned into 10 relatively equal-sized subsets and represented with subset numbers (i.e., 1, ... , 10). Each subset served exactly once as the source of validation samples. The models were trained using the monitoring data from the remaining nine subsets and then used to make predictions for the validation samples. The process was repeated 10 times so that each measured concentration had a paired predicted concentration. The predictive performance was measured with the coefficient of determination (R^2) and root mean square error (*RMSE*).

$$R^2 = 1 - \frac{\sum_{i=1}^{m} [y_i - \hat{y}_i]^2}{(y_i - \overline{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \quad (10)$$

where y_i is the measured concentrations of validation sample i ($i = 1, \cdots, m$), m is the number of validation samples, $\overline{y}$ is the mean of y_i, and $\hat{y}_i$ is the predicted concentrations.

4. Experimental Results

According to the modeling procedure in Section 3, we established the hybrid model and verified its prediction accuracy, as described in Section 4.1. Then, we utilized the hybrid model to obtain the 1 km monthly GOCs in Section 4.2.

4.1. Model Validation

As shown in the scatterplots of ozone observations (x-axis) against predicted values (y-axis) for each model (Figure 6), the hybrid model exhibits higher predictive performance (R^2 = 0.67, $RMSE$ = 15.87 μg/m^3) than the GLM (R^2 = 0.57, $RMSE$ = 17.88 μg/m^3) and BME model (R^2 = 0.65, $RMSE$ = 15.96 μg/m^3). In addition, for locations more than 100 km away from the nearest station, we compared the prediction accuracy of the hybrid model and the BME model (Figure 7). As shown in Figure 7, the hybrid model (R^2 = 0.53, $RMSE$ = 19.08 μg/m^3) has better predictive ability than the BME model (R^2 = 0.45, $RMSE$ = 20.71 μg/m^3) in areas without extensive monitoring sites.

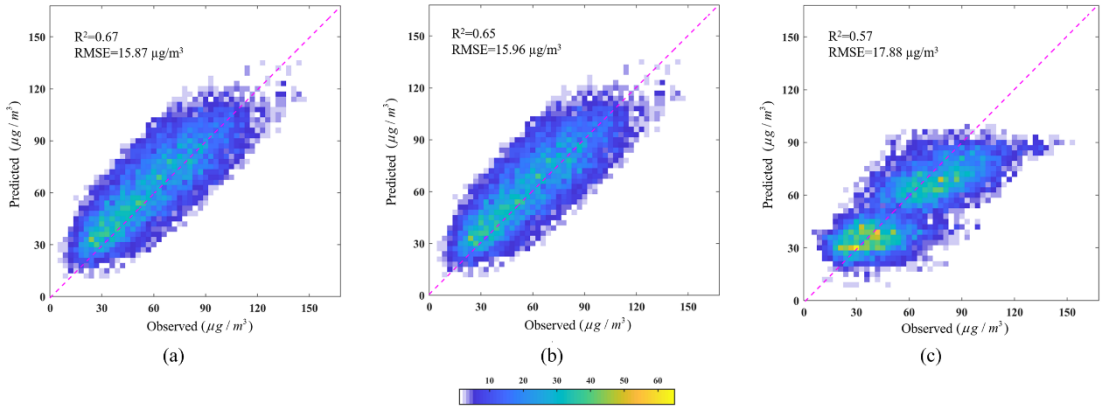

Figure 6. Density scatter plots of measurements and predictions for (**a**) hybrid model, (**b**) BME model, and (**c**) GLM (color bar represents the frequency of points).

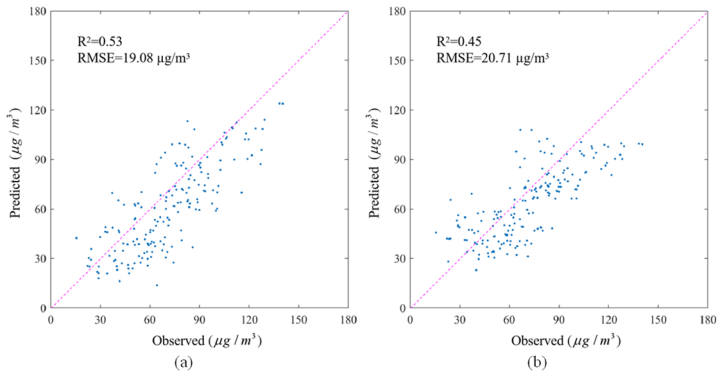

Figure 7. Scatter plots for (**a**) hybrid model and (**b**) BME model at locations more than 100 km from the nearest monitoring site.

We also evaluated the predictive performance of the hybrid model for different seasons (Figure 8) and geographical regions (Figure 9). The seasons were defined as spring (March–May), summer (June–August), autumn (September–November), and winter (December–February). As shown in Figure 8, the model performance showed seasonal variation, with a lower value of R^2 in winter and a higher value in summer. This result is consistent with those reported by Mo et al. (2021) [35], who predicted nationwide ground-level ozone in China.

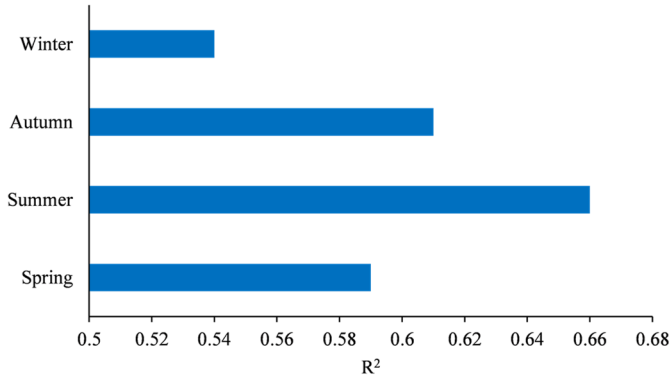

Figure 8. R^2-based accuracy evaluation of the hybrid model for different seasons.

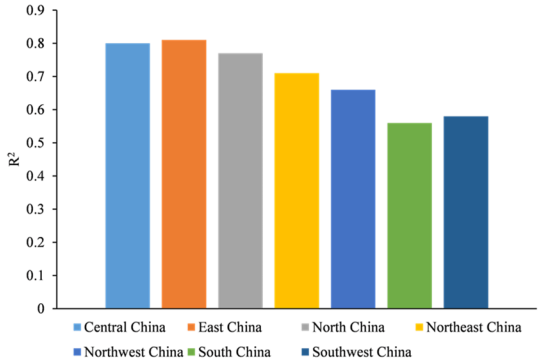

Figure 9. R^2-based accuracy evaluation of the hybrid model for different geographical regions.

As shown in Figure 9, the model performance exhibited spatial variations, with R^2 between 0.56 and 0.81. The best performance was achieved in East, Central, and North China, with R^2 values of 0.81, 0.80, and 0.77, respectively. A similar spatial pattern has also been reported in the literature [9,30,35]. The spatial variation could be affected by the sampling density and terrain condition. Dense monitoring stations improved the prediction accuracy of the hybrid model in East and Central China, while sparse monitoring stations reduced the prediction accuracy in Southwest China. It has been acknowledged that topographic effects have impacts on air convection, winds, precipitation and aerosol components [9], which in turn affect the predictive performance of the hybrid model. The relatively flat terrain in North and Northeast China resulted in higher predictive accuracy than the hilly and plateau landforms in the western part of China. Given the higher spatial resolution (i.e., 1 km $\times$ 1 km) and greater precision, the hybrid model is highly recommended for predicting GOCs at the national scale. Thus, we used the hybrid model to predict the 1 km monthly average GOCs across China.

4.2. Mapping of GOCs across China

Figure 10 illustrates the nationwide monthly average GOCs at a 1 km spatial resolution for China in 2018 predicted by the hybrid model. China has a vast territory with differences in population density, economic development level, industrialization level, topography, and other aspects in various regions, resulting in spatiotemporal heterogeneity in GOCs [42,43]. Temporally, the monthly average GOCs in most parts of China showed consistent change trajectories during the study period. They exhibited an increasing trend from January to May, which was followed by persistently high levels from May to August and then a decreasing trend from August to December. Summer is the most polluted season. According to previous studies, this is mainly attributed to seasonal variations in meteorological conditions; for instance, strong solar radiation and high temperatures lead to high concentrations in summer, while less sunlight and lower temperatures inhibit ozone formation in winter [9]. Spatially, the monthly average GOCs were higher in North,

East, and Northwest China and lower in Southwest China. North and East China are the population centers of China and have well-developed industries and agriculture. The large amounts of VOCs emitted from some petrochemical and organic industries and artificial sources resulted in their high GOCs. Northwest China is characterized by a high average elevation, strong solar radiation, low rainfall, and large temperature differences, which provide beneficial conditions for atmospheric photochemical reactions. The monthly average GOCs in Northwest China were particularly high in summer, while long winters with low temperatures inhibited ozone formation to some degree. Areas with relatively low monthly average GOCs were mainly located in Southwest China because local meteorological conditions (e.g., rainy and foggy, less sunlight and low temperature) and low precursor emissions were not favorable to ozone formation.

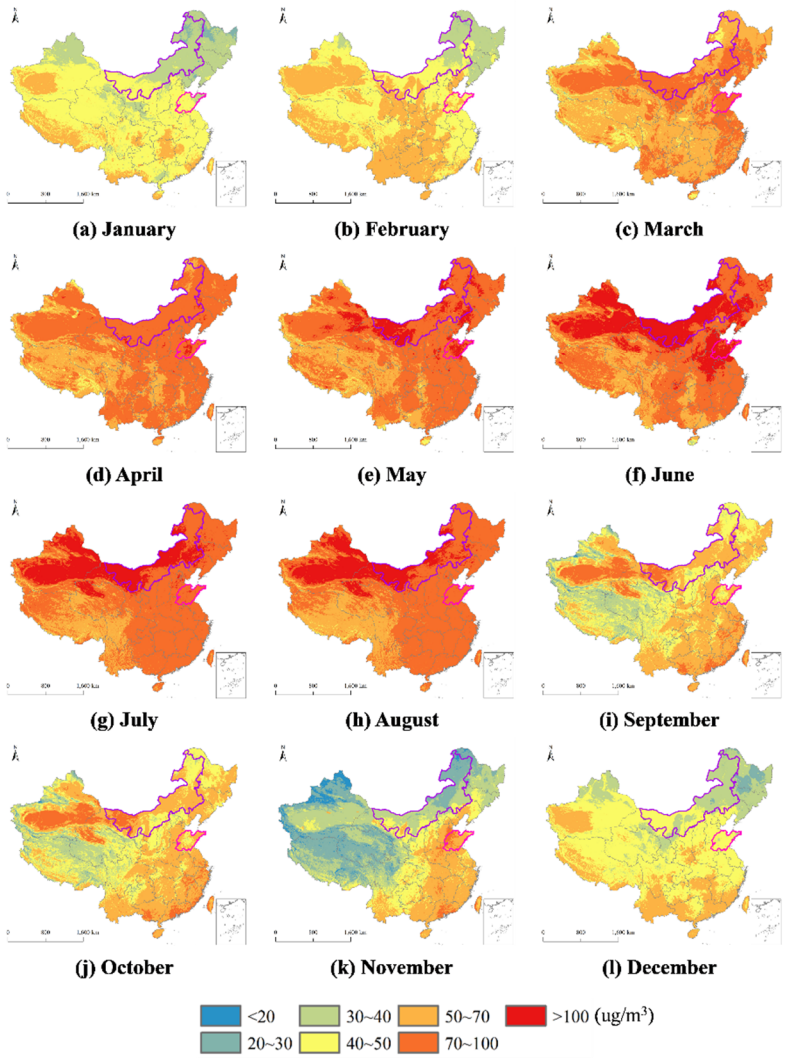

Figure 10. Nationwide monthly average GOCs at a 1 km spatial resolution in China in 2018 (purple box represents Inner Mongolia, pink box represents Shandong).

5. Discussion

In this study, a hybrid model that combines a GLM and a BME model was proposed to predict nationwide GOCs at a 1 km spatial resolution. The major findings of this paper are summarized in Section 5.1, the spatiotemporal heterogeneity of ozone distribution is illustrated in Section 5.2, the ozone exposure analysis based on the GOCs are discussed in Section 5.3, and the advantages and possible problems of the proposed model are described in Section 5.4.

5.1. Major Findings

There are two important findings in this paper. First, the fitted theoretical covariance model in Figure 4 indicates that the ground-level ozone measurements contributed significantly to the prediction of ≤111 km from the monitoring site. Second, our research demonstrates that the BME model was preferable for predicting GOCs within 100 km from a monitoring site, whereas the GLM performed better beyond that distance. Please kindly note that current studies [15] also argue for the inferior performance of BME in dealing with sparse monitoring sites, which is consistent with our result. On this basis, our work further presents the suggested distance, which can help colleagues choose suitable models regarding specific data accessibility.

5.2. Spatiotemporal Heterogeneity

The distribution of GOCs in China had spatiotemporal heterogeneity. Monthly average GOCs were higher in summer and lower in winter. The GOCs tend to form and accumulate to higher levels in the northern part of China. Hotspots located in Northwest and North China are attributed to local high temperatures, low rainfall, dusty weather, and ozone transport. Increased latitude led to higher concentrations in Northwest China than in Southwest China, but the concentrations in Southwest China were even higher in winter because the terrain obstructs ozone input from southbound westerly airflow. Highly populated areas, including the northern parts of Central and East China, show higher GOCs than other areas in the considered range.

5.3. Ozone Exposure Analysis Based on the GOCs

Ground-level ozone pollution induces numerous adverse human health problems. According to the air quality guideline proposed by the World Health Organization (WHO), the GOCs above 100 μg/m^3 can be hazardous to public health [30]. Long-term exposure to high level of GOCs (i.e., >100 μg/m^3) not only causes cardiovascular and respiratory diseases but also increases morbidity and mortality from chronic obstructive pulmonary disease (COPD) [34,44,45].

On the basis of the predicted GOCs provided by the hybrid model in this paper, we can further analyze the spatiotemporal distribution of ozone exposure for alleviating or even avoiding the human health hazards caused by ground-level ozone pollution. Figure 10 reveals the following:

(a) The monthly average GOCs tended to be higher than 100 μg/m^3 from May to August in 2018.
(b) About 28% of the Chinese population lived in areas (mainly distributed in Northeast, North, and Northwest China) with monthly average GOCs higher than 100 μg/m^3 during the above periods.

Using the predicted GOCs of high spatial resolution (i.e., 1 km × 1 km) provided by the hybrid model, ozone exposure at provincial scale or below (e.g., county scale and village scale) can be obtained. Taking the provincial scale as an example, Inner Mongolia (i.e., the purple box in Figure 10) as one of the severe pollution provinces has higher monthly average GOCs than other provinces in summer, which is associated to the high temperatures, low rainfall, high altitude, dusty weather, and ozone transport. Shandong (i.e., the pink box in Figure 10) as a large populous province (accounts for about 7.2% of the Chinese population) with well-developed industries and agriculture has 5 months (from April to August) with monthly average GOCs higher than 100 μg/m^3, and this is attributed to the large amounts of ozone precursors emitted from some petrochemical, heavy industries, and artificial sources [46–48].

For the megacity cluster region in China, the Beijing–Tianjin–Hebei region (BTH) has the most months (about 5 months) with monthly average GOCs higher than 100 μg/m^3. The high GOCs in BTH may be caused by the emissions of precursor pollutants and the meteorological conditions conducive to photochemical reaction [48]. In addition, because of the significantly increased rainfall in BTH in summer, NOx was easily converted to

nitrate under high temperature and humidity conditions, which reduced the efficiency of GOC removal by vapor-phase chemistry, thus leading to ozone pollution. In future work, we will investigate the prediction of long-term concentrations (e.g., historical ozone concentration beyond the model years) and concentrations with finer temporal resolution (e.g., daily level) to meet the requirement of exposure assessment.

5.4. Advantages and Possible Problems of the Proposed Model

The advantages of the hybrid model included the desirable retrieval accuracy (R^2 = 0.67, *RMSE* = 15.26 μg/m^3) at high spatial resolution and the superior generality for nationwide mapping. In particular, for locations more than 100 km away from the nearest station, the hybrid model (R^2 = 0.53, *RMSE* = 19.08 μg/m^3) outperformed the BME model (R^2 = 0.45, *RMSE* = 20.71 μg/m^3). In addition, the R^2 values of the hybrid model reached 0.81, 0.80, and 0.77 for East, Central, and North China, respectively. The hybrid model also exhibited satisfactory performance for the four seasons, especially summer (R^2 = 0.66).

The predictive performance of the large-scale and high-resolution hybrid model is superior to that of previous deterministic models and is comparable to that of current statistical models (Table 3). First, the hybrid model outperforms two Weather Research and Foresting (WRF)-Community Multiscale Air Quality (CMAQ) methods [34,49], indicating the superiority of statistical models. Second, the qualitative indices of the high spatial resolution (i.e., 1 km × 1 km) hybrid model are comparable to others' rough statistical models, which indicates that the proposed work can provide finer spatial information with comparable confidence. In more detail, the hybrid model yields a similar R^2 to previous nationwide studies conducted at a 0.1-degree spatial resolution using the eXtreme Gradient Boosting algorithm (R^2 ranged from 0.60 to 0.87) [19] and random forest method (R^2 = 0.71 and *RMSE* = 19 μg/m^3) [30]. Thus, the hybrid model is highly recommended for predicting GOCs at the national scale.

Table 3. Comparison with other's results.

	Spatial Resolution	R^2	*RMSE*
Liu et al. (2018)	>0.1°	>0.6	-
Lin et al. (2018)	36 km	>0.5	-
Liu et al. (2020)	0.1°	0.60 to 0.87	-
Zhan et al. (2018)	0.1°	0.71	19 μg/m^3
Proposed hybrid model	1 km	0.67	15.26 μg/m^3

The main limitation of the hybrid model originated from the seven explanatory variables utilized, which have been shown to partially explain the complex formation of ground-level ozone. Please note that collecting relevant explanatory variables for large-scale mapping is difficult and time-consuming. Therefore, we appeal to relevant authorities to open more data, which will contribute to the prediction of high-resolution GOCs over wide areas.

6. Conclusions

To predict nationwide monthly average GOCs at a 1 km spatial resolution for China in 2018, a hybrid model that combines a GLM and a BME model was developed based on a specific broad range of each submodel. The hybrid model was constructed based on the BME model for measurements within 100 km from their nearest monitoring sites and GLM when the distance was beyond 100 km. The hybrid model exhibited satisfactory performance (R^2 = 0.67, *RMSE* = 15.26 μg/m^3) in predicting nationwide GOCs at high spatial resolution (i.e., 1 km). Although the derived nationwide spatiotemporal characteristics of GOCs may be uncertain in some areas, they offer valuable information on the spatiotemporal patterns of GOCs on the national scale, which is helpful for ground-level ozone pollution control and prevention in China.

Author Contributions: Y.M. proposed the study, designed and carried out data analyses. Y.M. and J.L. contributed to manuscript writing and revision. D.X. contributed to acquisition and analyses of sample data. J.Z. helped with reference data sources. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the National Social Science Foundation of China (grant number 19CSH004), the General Financial Grant from China Postdoctoral Science Foundation (grant number 2019M662723).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Javanmardi, P.; Morovati, P.; Farhadi, M.; Geravandi, S.; Khaniabadi, Y.O.; Angali, K.A.; Taiwo, A.M.; Sicard, P.; Goudarzi, G.; Valipour, A.; et al. Monitoring the impact of ambient ozone on human health using time series analysis and air quality model approaches. *Fresenius Environ. Bull.* **2018**, *27*, 533–544.
2. Sicard, P.; Anav, A.; De Marco, A.; Paoletti, E. Projected global ground-level ozone impacts on vegetation under different emission and climate scenarios. *Atmos. Chem. Phys.* **2017**, *17*, 12177–12196. [CrossRef]
3. Chattopadhyay, S.; Chattopadhyay, G. Modeling and Prediction of Monthly Total Ozone Concentrations by Use of an Artificial Neural Network Based on Principal Component Analysis. *Pure Appl. Geophys.* **2012**, *169*, 1891–1908. [CrossRef]
4. Gao, M.; Gao, J.; Zhu, B.; Kumar, R.; Lu, X.; Song, S.; Zhang, Y.; Jia, B.; Wang, P.; Beig, G.; et al. Ozone pollution over China and India: Seasonality and sources. *Atmos. Chem. Phys.* **2020**, *20*, 4399–4414. [CrossRef]
5. Jaber, S.M.; Abu-Allaban, M.M. Mapping the spatial distribution of tropospheric ozone and exploring its association with elevation and land cover over North Jordan. *J. Spat. Sci.* **2017**, *62*, 307–322. [CrossRef]
6. Christakos, G.; Kolovos, A.; Serre, M.L.; Vukovich, F. Total ozone mapping by integrating databases from remote sensing instruments and empirical models. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 991–1008. [CrossRef]
7. Abdul-Wahab, S.A.; Al-Alawi, S.M. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environ. Model. Softw.* **2002**, *17*, 219–228. [CrossRef]
8. Castellano, M.; Franco, A.; Cartelle, D.; Febrero, M.; Roca, E. Identification of NOx and Ozone Episodes and Estimation of Ozone by Statistical Analysis. *Water Air Soil Pollut.* **2009**, *198*, 95–110. [CrossRef]
9. Cheng, L.; Wang, S.; Gong, Z.; Li, H.; Yang, Q.; Wang, Y. Regionalization based on spatial and seasonal variation in ground-level ozone concentrations across China. *J. Environ. Sci.* **2018**, *67*, 179–190. [CrossRef]
10. Grigoras, G.; Stefan, S.; Rada, C.; Grigoras, C. Assessing of surface-ozone concentration in Bucharest, Romania, using OML and satellite data. *Atmos. Pollut. Res.* **2016**, *7*, 567–576. [CrossRef]
11. Zunckel, M.; Koosailee, A.; Yarwood, G.; Maure, G.; Venjonoka, K.; Van Tienhoven, A.M.; Otter, L. Modelled surface ozone over southern Africa during the Cross Border Air Pollution Impact Assessment Project. *Environ. Model. Softw.* **2006**, *21*, 911–924. [CrossRef]
12. Camalier, L.; Cox, W.; Dolwick, P. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmos. Environ.* **2007**, *41*, 7127–7137. [CrossRef]
13. Gong, X.; Kaulfus, A.; Nair, U.; Jaffe, D.A. Quantifying O3 Impacts in Urban Areas Due to Wildfires Using a Generalized Additive Model. *Environ. Sci. Technol.* **2017**, *51*, 13216–13223. [CrossRef] [PubMed]
14. Sun, W.; Zhang, H.; Palazoglu, A. Prediction of 8 h-average ozone concentration using a supervised hidden Markov model combined with generalized linear models. *Atmos. Environ.* **2013**, *81*, 199–208. [CrossRef]
15. Adam-Poupart, A.; Brand, A.; Fournier, M.; Jerrett, M.; Smargiassi, A. Spatiotemporal Modeling of Ozone Levels in Quebec (Canada): A Comparison of Kriging, Land-Use Regression (LUR), and Combined Bayesian Maximum Entropy–LUR Approaches. *Environ. Health Perspect.* **2014**, *122*, 970–976. [CrossRef]
16. Bogaert, P.; Christakos, G.; Jerrett, M.; Yu, H.-L. Spatiotemporal modelling of ozone distribution in the State of California. *Atmos. Environ.* **2009**, *43*, 2471–2480. [CrossRef]
17. Ren, X.; Mi, Z.; Georgopoulos, P.G. 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. *Environ. Int.* **2020**, *142*, 105827. [CrossRef]
18. Biancofiore, F.; Verdecchia, M.; Di Carlo, P.; Tomassetti, B.; Aruffo, E.; Busilacchio, M.; Bianco, S.; Di Tommaso, S.; Colangeli, C. Analysis of surface ozone using a recurrent neural network. *Sci. Total Environ.* **2015**, *514*, 379–387. [CrossRef] [PubMed]
19. Liu, R.; Ma, Z.; Liu, Y.; Shao, Y.; Zhao, W.; Bi, J. Spatiotemporal distributions of surface ozone levels in China from 2005 to 2017: A machine learning approach. *Environ. Int.* **2020**, *142*, 105823. [CrossRef]
20. Chaloulakou, A.; Saisana, M.; Spyrellis, N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Sci. Total Environ.* **2003**, *313*, 1–13. [CrossRef]
21. Spellman, G. An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom. *Appl. Geogr.* **1999**, *19*, 123–136. [CrossRef]
22. Gorai, A.K.; Tchounwou, P.B.; Mitra, G. Spatial Variation of Ground Level Ozone Concentrations and its Health Impacts in an Urban Area in India. *Aerosol Air Qual. Res.* **2017**, *17*, 951–964. [CrossRef] [PubMed]

23. Solaiman, T.A.; Coulibaly, P.; Kanaroglou, P. Ground-level ozone forecasting using data-driven methods. *Air Qual. Atmos. Health* **2008**, *1*, 179–193. [CrossRef]
24. Yang, J.; Wen, Y.; Wang, Y.; Zhang, S.; Pinto, J.P.; Pennington, E.A.; Wang, Z.; Wu, Y.; Sander, S.P.; Jiang, J.H.; et al. From COVID-19 to future electrification: Assessing traffic impacts on air quality by a machine-learning model. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2102705118. [CrossRef]
25. Pak, U.; Kim, C.; Ryu, U.; Sok, K.; Pak, S. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. *Air Qual. Atmos. Health* **2018**, *11*, 883–895. [CrossRef]
26. Lu, X.; Hong, J.; Zhang, L.; Cooper, O.R.; Schultz, M.G.; Xu, X.; Wang, E.T.; Galo, M.; Zhao, Y.; Zhang, Y. Severe Surface Ozone Pollution in China: A Global Perspective. *Environ. Sci. Technol. Lett.* **2018**, *5*, 487–494. [CrossRef]
27. Wang, X.; Manning, W.; Feng, Z.; Zhu, Y. Ground-level ozone in China: Distribution and effects on crop yields. *Environ. Pollut.* **2007**, *147*, 394–400. [CrossRef] [PubMed]
28. Ozbay, B.; Keskin, G.A.; Dogruparmak, S.C.; Ayberk, S. Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models. *Ecol. Inform.* **2011**, *6*, 242–247. [CrossRef]
29. Zhao, H.; Zheng, Y.; Li, T.; Wei, L.; Guan, Q. Temporal and Spatial Variation in, and Population Exposure to, Summertime Ground-Level Ozone in Beijing. *Int. J. Environ. Res. Public Health* **2018**, *15*, 628. [CrossRef]
30. 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]
31. Al-Alawi, S.M.; Abdul-Wahab, S.A.; Bakheit, C.S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environ. Model. Softw.* **2008**, *23*, 396–403. [CrossRef]
32. Zheng, J.; Swall, J.L.; Cox, W.M.; Davis, J.M. Interannual variation in meteorologically adjusted ozone levels in the eastern United States: A comparison of two approaches. *Atmos. Environ.* **2007**, *41*, 705–716. [CrossRef]
33. Chen, L.; Gao, S.; Zhang, H.; Sun, Y.; Ma, Z.; Vedal, S.; Mao, J.; Bai, Z. Spatiotemporal modeling of PM 2.5 concentrations at the national scale combining land use regression and Bayesian maximum entropy in China. *Environ. Int.* **2018**, *116*, 300–307. [CrossRef]
34. 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]
35. Mo, Y.; Li, Q.; Karimian, H.; Zhang, S.; Kong, X.; Fang, S.; Tang, B. Daily spatiotemporal prediction of surface ozone at the national level in China: An improvement of CAMS ozone product. *Atmos. Pollut. Res.* **2021**, *12*, 391–402. [CrossRef]
36. Duan, S.-B.; Li, Z.-L.; Leng, P. A framework for the retrieval of all-weather land surface temperature at a high spatial resolution from polar-orbiting thermal infrared and passive microwave data. *Remote Sens. Environ.* **2017**, *195*, 107–117. [CrossRef]
37. Gelaro, R.; McCarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.A.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef] [PubMed]
38. Yu, H.-L.; Kolovos, A.; Christakos, G.; Chen, J.-C.; Warmerdam, S.; Dev, B. Interactive spatiotemporal modelling of health systems: The SEKS–GUI framework. *Stoch. Environ. Res. Risk Assess.* **2007**, *21*, 555–572. [CrossRef]
39. Yang, Y.; Christakos, G. Spatiotemporal Characterization of Ambient PM2.5 Concentrations in Shandong Province (China). *Environ. Sci. Technol.* **2015**, *49*, 13431–13438. [CrossRef]
40. Rao, S.T.; Zurbenko, I.G.; Neagu, R.; Porter, P.S.; Ku, J.Y.; Henry, R.F. Space and time scales in ambient ozone data. *Bull. Am. Meteorol. Soc.* **1997**, *78*, 2153–2166. [CrossRef]
41. Lee, S.J.; Serre, M.L.; van Donkelaar, A.; Martin, R.V.; Burnett, R.T.; Jerrett, M. Comparison of Geostatistical Interpolation and Remote Sensing Techniques for Estimating Long-Term Exposure to Ambient PM(2.5) Concentrations across the Continental United States. *Environ. Health Perspect.* **2012**, *120*, 1727–1732. [CrossRef] [PubMed]
42. Wang, Y.; Zhang, Y.; Hao, J.; Luo, M. Seasonal and spatial variability of surface ozone over China: Contributions from background and domestic pollution. *Atmos. Chem. Phys.* **2011**, *11*, 3511–3525. [CrossRef]
43. Yin, C.; Solmon, F.; Deng, X.; Zou, Y.; Deng, T.; Wang, N.; Li, F.; Mai, B.; Liu, L. Geographical distribution of ozone seasonality over China. *Sci. Total Environ.* **2019**, *689*, 625–633. [CrossRef] [PubMed]
44. Anenberg, S.C.; Horowitz, L.W.; Tong, D.Q.; West, J.J. An Estimate of the Global Burden of Anthropogenic Ozone and Fine Particulate Matter on Premature Human Mortality Using Atmospheric Modeling. *Environ. Health Perspect.* **2010**, *118*, 1189–1195. [CrossRef]
45. Levy, J.I.; Chemerynski, S.M.; Sarnat, J.A. Ozone exposure and mortality: An empiric Bayes metaregression analysis. *Epidemiology* **2005**, *16*, 458–468. [CrossRef]
46. Li, L.; An, J.; Shi, Y.; Zhou, M.; Yan, R.; Huang, C.; Wang, H.; Lou, S.; Wang, Q.; Lu, Q.; et al. Source apportionment of surface ozone in the Yangtze River Delta, China in the summer of 2013. *Atmos. Environ.* **2016**, *144*, 194–207. [CrossRef]
47. Wu, R.R.; Xie, S.D. Spatial Distribution of Ozone Formation in China Derived from Emissions of Speciated Volatile Organic Compounds. *Environ. Sci. Technol.* **2017**, *51*, 2574–2583. [CrossRef] [PubMed]
48. Yang, G.; Liu, Y.; Li, X. Spatiotemporal distribution of ground-level ozone in China at a city level. *Sci. Rep.* **2020**, *10*, 1–12. [CrossRef]
49. Lin, Y.; Jiang, F.; Zhao, J.; Zhu, G.; He, X.; Ma, X.; Li, S.; Sabel, C.E.; Wang, H. Impacts of O-3 on premature mortality and crop yield loss across China. *Atmos. Environ.* **2018**, *194*, 41–47. [CrossRef]