

## Article

# Haze Risk Assessment Based on Improved PCA-MEE and ISPO-LightGBM Model

Hongbin Dai <sup>1</sup>, Guangqiu Huang <sup>1</sup>, Huibin Zeng <sup>1</sup> and Rongchuan Yu <sup>2,\*</sup><sup>1</sup> School of Management, Xi'an University of Architecture and Technology, Xi'an 710055, China<sup>2</sup> School of Mathematics and Computer Science, Guangxi Science & Technology of Normal University, Laibin 546199, China

\* Correspondence: yurongchuan@gxstnu.edu.cn; Tel.: +86-152-7710-7077

**Abstract:** With the economic development in China, haze risks are frequent. It is important to study the urban haze risk assessment to manage the haze disaster. The haze risk assessment indexes of 11 cities in Fenwei Plain were selected from three aspects: the sensitivity of disaster-inducing environments, haze component hazards and the vulnerability of disaster-bearing bodies, combined with regional disaster system theory. The haze hazard risk levels of 11 cities in Fenwei Plain were evaluated using the matter-element extension (MEE) model, and the indicator weights were determined by improving the principal component analysis (PCA) method using the entropy weight method, and finally, five haze hazard risk assessment models were established by improving the particle swarm optimization (IPSO) light gradient boosting machine (LightGBM) algorithm. It is used to assess the risk of affected populations, transportation damage risk, crop damage area risk, direct economic loss risk and comprehensive disaster risk before a disaster event occurs. The experimental comparison shows that the haze risk index of Xi'an city is the highest, and the full index can improve the evaluation accuracy by 4–16% compared with only the causative factor index, which indicates that the proposed PCA-MEE-ISPO-LightGBM model evaluation results are more realistic and reliable.

**Keywords:** haze hazard; principal component analysis; entropy weight method; matter-element extension model; risk assessment; ISPO-LightGBM

**Citation:** Dai, H.; Huang, G.; Zeng, H.; Yu, R. Haze Risk Assessment Based on Improved PCA-MEE and ISPO-LightGBM Model. *Systems* **2022**, *10*, 263. <https://doi.org/10.3390/systems10060263>

Academic Editors: Baojie He, Linchuan Yang and Junqing Tang

Received: 17 November 2022

Accepted: 16 December 2022

Published: 19 December 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 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

Air pollution in 2017 was estimated to have caused about 4.9 million deaths globally, while PM<sub>2.5</sub> alone was responsible for 2.94 million deaths [1]. In China, the WHO's annual median PM<sub>2.5</sub> concentration model shows that only parts of Tibet meet the organization's air quality guidelines [2]. The severely hazardous haze also causes between 1.2 and 1.6 million premature deaths per year [3]. The excessive consumption of fossil fuels, such as coal, has been shown in previous studies to be responsible for significant deteriorations in air quality [4,5]. The contribution of the secondary aerosol formation of VOCs to haze formation is significant [6,7]. VOCs are a key precursor for the formation of O<sub>3</sub> and secondary organic aerosols (SOA) [8]. SOA are an important component of fine particulate matter and a major contributor to haze pollution [9,10]. Studies have shown that haze pollution in China is mainly driven by SOA [11]. Severe haze pollution leads to poor air quality and an estimated 2.6 to 4.8 million premature deaths worldwide each year [12–14]. There is direct evidence of the human health effects of haze air pollution exposure related to respiratory VOC biomarkers, such as propanol and isoprene, in haze pollution [15].

How to prevent and control urban haze disasters has become one of the major issues facing China's sustainable economic development and harmonious urban development. The formation of urban haze has many causal factors, a wide impact, and a social and

complex nature. The risk components of urban haze vary from place to place, and the study of urban haze risk assessment is important for proposing countermeasures for urban haze management from the perspective of risk prevention and control. This paper is expected to provide decision-making references for the development of targeted urban disaster prevention and mitigation systems, energy conservation and environmental protection, and haze management.

Haze events are frequent in China's highly industrialized, economically developed and densely populated urban areas with a long duration and record air pollutant concentrations [16]. Urbanization has provided inexhaustible impetus for China's economic development, which raises the question of the relationship between the advancement of urbanization and haze pollution. Some scholars have used the development of urbanization as the main research variable through the construction of mathematical models to show that the development of urbanization exacerbates haze pollution [17,18]. Singh et al. also showed that for most PM<sub>2.5</sub> haze-causing studies in South Asia, vehicle emissions emerged as the dominant source [19]. Latif et al. found that local vehicle emissions and industrial activities are significant contributors to haze pollutants in Malaysia [20].

Researchers have studied the atmospheric haze causality of haze systems using a variety of methods, such as Zhang et al. who combined causal analysis and stochastic nonlinear features to construct a haze hazard prediction model for Beijing and simulated haze hazard trends under different governance and control policies [21]. Several researchers have studied mathematical models for haze prediction, including nonparametric regression models [22,23], deep recurrent neural networks [24,25], inverse matrix-free machine learning models [26], the nonlinear gray model [27] and graphical networks [28,29]. These methods avoid the analysis of the complex details and mechanisms of haze hazards. Clarifying the causal relationships among the factors influencing haze hazards is a prerequisite for building haze prediction models. To explore the atmospheric haze causality of haze systems, some other methods have been applied including Granger causality analysis [30,31], convergent cross mapping [32] and machine learning [33,34]. Factors influencing the haze hazard include oceanic transport at the marine level [35,36], local and global pollution emissions [37,38] and the interaction of industrial emissions with atmospheric dispersion [39,40].

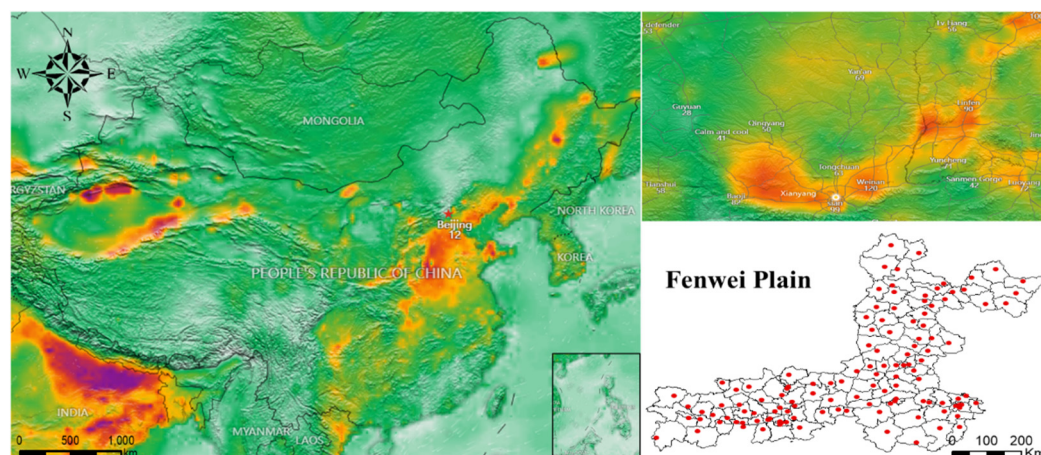
Unfortunately, these studies were unable to describe the dynamic formation and evolutionary mechanisms of cross-regional haze hazards. The above methods have improved the efficiency of the assessment to a certain extent, but there are still obvious shortcomings, mainly because it is not easy to explain the role of each model parameter, which is similar to a "black box" operation and cannot explain the role of different indicators in the disaster risk assessment. Meanwhile, the original risk assessment model of urban haze pollution loss is slow in conducting risk assessments and has the problem of poor accuracy of the assessment results. In this paper, we selected the haze disaster cases in Fenwei Plain of China as training samples, collected 13 indicators that may affect the haze disaster risk at a county level and established a haze disaster risk assessment process model based on the PCA-MEE-ISPO-LightGBM algorithm.

## 2. Data Sources and Methods

### 2.1. Data Sources

Economic density and population density are from the Data Center for Resource and Environmental Sciences, Chinese Academy of Sciences, 2016–2021 (<https://www.resdc.cn/Default.aspx>) (accessed on 17 November 2022). The annual average concentrations of PM<sub>10</sub>, PM<sub>2.5</sub>, SO<sub>2</sub>, VOCs and NO<sub>2</sub> for 2016–2021 were calculated from the daily data downloaded from the National Real-Time Urban Air Quality Release Platform (<http://106.37.208.233:20035/>) (accessed on 17 November 2022). Other data are mainly from the *Shaanxi Provincial Statistical Yearbook 2016–2021*, the *Henan Provincial Statistical Yearbook 2016–2021*, the *Shanxi Provincial Statistical Yearbook 2016–2021* and the statistical yearbooks

of prefecture-level cities. As shown in Figure 1, this paper selects the Fenwei Plain as the study area including 11 cities in Shaanxi, Shanxi and Henan. Red dots indicate 129 counties' data.



**Figure 1.** Location of study area.

According to the regional hazard system theory [41], in the formation of a disaster, the risk of disastrous factors, environmental sensitivity of the disaster and the disaster-bearing body are indispensable. The environmental sensitivity of the disaster refers to the earth's surface environment, including the natural and man-made environment, where the disaster-causing factors are formed in the disaster-causing environment and directly lead to the occurrence of the disaster. The disaster-bearing body refers to the object that suffers from the disaster and is adversely affected. The risk of disastrous factors, environmental sensitivity of the disaster and disaster-bearing body jointly determine the magnitude of the haze disaster risk. Research shows that the main material components of urban haze are toxic gases and respirable particulate matter, which mainly come from human production and life; the formation of urban haze is influenced by human factors. In terms of the anthropogenic factors, the more developed the economy, the more motor vehicles owned by urban residents, the more exhaust emissions from motor vehicles and the increased risk of haze disasters. The development of the secondary industry is often accompanied by pollution and damage to the environment. The more a region relies on the secondary industry for its economic development, the more serious the pollution and damage to the environment and the greater the sensitivity to haze disasters. The greater the consumption of coal in a region, the more industrial emissions and the greater the risk of haze disaster. In addition, building construction is also a major source of respirable particulate matter; the larger the area of building construction in a region, the higher the concentration of respirable particulate matter and the greater the risk of haze disaster.

The regional hazard system is an earth surface heterogeneous system composed of the three aforementioned factors, and the disaster risk is influenced by the combined effect of the three aforementioned factors. In this paper, five indicators are selected from anthropogenic factors to quantitatively evaluate the sensitivity of the environmental sensitivity of disaster in the Fenwei Plain, including economic density, which represents the degree of economic development in a city; the number of motor vehicles, which represents the amount of motor vehicle emissions; the share of secondary industry, which represents the dependence of a region's economy on the secondary industry; the share of coal consumption, which represents the industrial pollution emissions of a city; and the area of housing construction, which represents the housing construction projects of a region. The harmfulness of haze components refers to the damage of various components of haze to urban economy and residents' health. As shown in Table 1, five indicators, including the annual average concentration of VOCs, PM<sub>10</sub>, PM<sub>2.5</sub>, SO<sub>2</sub> and NO<sub>2</sub> were selected as the evaluation

indicators of haze component vulnerability in Fenwei Plain. The population density, the number of health institutions and green areas in built-up areas are selected as the evaluation indicators of the vulnerability of urban haze. In hazy weather, the concentration of aerosols in the air rises, the atmospheric layer is relatively stable and unfavorable for the convective diffusion of air, the humidity and visibility of the atmosphere change dramatically, and people's lives and health are greatly adversely affected. The greater the population density of a city, the greater the number of people suffering from haze and the greater the vulnerability of the disaster-bearing body. The greater the number of health institutions, the more developed the medical care, and the higher the carrying capacity of the medical system in the area, the more people can be treated and cured from the haze, reducing the risk of the haze and the vulnerability of the city to haze. Urban greening can absorb harmful gases and dust, reduce air pollution and reduce the vulnerability of the disaster-bearing body. Therefore, the population density, number of health institutions and green areas in built-up areas are selected as the evaluation indexes of urban haze vulnerability.

**Table 1.** Fenwei Plain haze disaster risk assessment index system.

First-Grade Indexes	Second-Grade Indexes	Unit	Abbreviations	Positive/Negative Indicators
Risk of disastrous factors (RD)	Economic density	CNY Billion/km <sup>2</sup>	ED	Positive indicators
	Number of motor vehicles	Num.	NMV	Positive indicators
	Percentage of secondary industry	%	SSVP	Positive indicators
	House construction area	m <sup>2</sup>	HCA	Positive indicators
	Share of coal consumption	%	COC	Positive indicators
Environmental sensitivity of disaster (ESD)	VOCs	µg/m <sup>3</sup>	-	Positive indicators
	PM <sub>10</sub>	µg/m <sup>3</sup>	-	Positive indicators
	PM <sub>2.5</sub>	µg/m <sup>3</sup>	-	Positive indicators
	SO <sub>2</sub>	µg/m <sup>3</sup>	-	Positive indicators
	NO <sub>2</sub>	µg/m <sup>3</sup>	-	Positive indicators
Disaster-bearing body (DBB)	Population density	Persons/km <sup>2</sup>	PD	Positive indicators
	Number of health institutions	Num.	NHI	Negative indicators
	Greening area of built-up area	hm <sup>2</sup>	ACB	Negative indicators

## 2.2. Methods

### 2.2.1. Matter-Element Extension Model

The basic idea of the matter-element extension model is to first delineate the categories of objects to be evaluated and delineate the different categories according to the relevant research results. Extension is a subject based on extension mathematics and matter element theory; matter element is the logical cell of extension [42]. Assuming that the name of the thing is  $N$ , the response thing feature is  $C$ , and the value range of  $C$  is  $V$ , the ordered triple  $R = \{N, C, V\}$  can be used as the basic matter element to describe the thing. The risk of haze disaster caused by VOCs is defined as the basic matter element  $R$ , then  $N$  represents the risk of haze disaster,  $C$  represents the risk characteristics, and  $V$  is the characteristic value. If  $N$  has  $n$  features  $C_1, C_2, \dots, C_n$ , then

$$R = \begin{pmatrix} N & C_1 & V_1 \\ & C_2 & V_2 \\ & \vdots & \vdots \\ & C_n & V_n \end{pmatrix} \quad (1)$$

(1) Determine the classical domain

According to the risk of haze disaster caused by VOCs, the risk assessment of haze disaster caused by VOCs is divided into  $e$  classification levels ( $e = 1, 2, \dots, s$ ). The risk level of haze disaster is set, and  $C_j$  is the evaluation index of emergency management ability ( $j = 1, 2, \dots, n$ ). The value range of  $C_j$  is  $V_{0ej}$ , and its classical domain can be expressed as

$$R_{0e} = (N_{0e}, C_j, V_{0ej}) \begin{pmatrix} N & C_1 & \langle a_{0e1}, b_{0e1} \rangle \\ & C_2 & \langle a_{0e2}, b_{0e2} \rangle \\ & \vdots & \vdots \\ & C_n & \langle a_{0en}, b_{0en} \rangle \end{pmatrix} \quad (2)$$

$R_{0e}$  is the classical domain matter element.

(2) Determine the nodal domain

The matter-element of VOC-induced haze disaster risk assessment is essentially the atmosphere corresponding to each evaluation index (the range from the lowest value to the highest value). The eigenvalues of the object unit  $N_p$  ( $p = 1, 2, \dots, m$ ) can be evaluated according to the actual situation and scored according to the classification criteria, establishing the matter-element to be evaluated. The matter-element to be evaluated can be expressed as:

$$R_p = (N_p, C_j, V_{pj}) \begin{pmatrix} N_p & C_1 & \langle a_{p1}, b_{p1} \rangle \\ & C_2 & \langle a_{p2}, b_{p2} \rangle \\ & \vdots & \vdots \\ & C_n & \langle a_{pn}, b_{pn} \rangle \end{pmatrix} \quad (3)$$

$R_p$  is a nodal matter element;  $n_p$  is the individual to be tested for VOC-induced haze disaster risk assessment; and  $V_{pj} = \langle a_{pn}, b_{pn} \rangle$  is the magnitude range of the node domain matter element with respect to the characteristic  $C_j$ , where  $\langle a_{0ej}, b_{0ej} \rangle \dots \langle a_{pj}, b_{pj} \rangle$ , ( $j = 1, 2, \dots, n$ ).

(3) Establish evaluation index correlation function

Calculating the correlation coefficient of evaluation index, the correlation function is:

$$k_e(v_i) = \begin{cases} \frac{\rho[v_i(e), v_{0ej}]}{\rho[v_i(e), v_{pj}] - \rho[v_i(e), v_{0ej}]}, & (\rho[v_i(e), v_{pj}] - \rho[v_i(e), v_{0ej}] \neq 0) \\ -\rho[v_i(e), v_{0ej}] - 1, & (\rho[v_i(e), v_{pj}] - \rho[v_i(e), v_{0ej}] = 0) \end{cases} \quad (4)$$

$$\begin{aligned} \rho[v_i(e), V_{0ej}] &= \left| v_i - \frac{a_{0ej} + b_{0ej}}{2} \right| - \frac{b_{0ej} - a_{0ej}}{2} \\ \rho[v_i(e), V_{pj}] &= \left| v_i - \frac{a_{pj} + b_{pj}}{2} \right| - \frac{b_{pj} - a_{pj}}{2} \end{aligned} \quad (5)$$

(4) Define the entropy of evaluation indicators

$$f_{ki} = \frac{r_{ki}}{\sum_{k=1}^m r_{ki}} \quad (6)$$

$r_{ki}$  denotes the element of the  $k$  th row and  $i$  th column of the normalized matrix. Let  $f_{ki}$  denote the element of the  $k$  th row and  $i$  th column of the evaluation index after standardization.

Then, the entropy  $H_i$  of the evaluation index is

$$H_i = \frac{-1}{\ln m} \sum_{k=1}^m f_{ki} \ln f_{ki} \quad (7)$$

where the constant  $\frac{-1}{\ln m}$  denotes the information entropy coefficient,

$$\lim_{f_{ki} \rightarrow 0} f_{ki} \ln f_{ki} = 0, 0 \leq H_i \leq 1.$$

Determine the entropy weight of each evaluation index  $w_i$ . Calculate the entropy weight of the evaluation index using the following formula:

$$w_i = \frac{1 - H_i}{n - \sum_{i=1}^n H_i} \quad \text{and} \quad \sum_{i=1}^n w_i = 1. \quad (8)$$

(5) Calculate the comprehensive correlation degree of each evaluation index

The comprehensive correlation degree of each evaluation index, also known as multi-factor comprehensive correlation degree, refers to the degree of belonging of the evaluation index to each evaluation grade, which can be expressed as:

$$k_e(N) = \sum_{i=1}^m w_i k_e(v_i) \quad (9)$$

In the formula,  $w_i$  is the weight vector of each evaluation index and satisfies

$$\sum_{i=1}^m w_i = 1.$$

(6) Calculate the VOCs' haze disaster risk assessment level

The general matter-element extension model criterion adopts the principle of maximum membership degree; that is,  $k_{ie0}(N) = \max \{k_{ie}(N) | e=1,2,\dots,s\}$ . Then, the risk level  $N$  of VOCs haze disaster to be evaluated belongs to level  $e$ . This method sometimes cannot contain complete evaluation information. The use of asymmetric closeness principle can better solve the problem of maximum membership principle failure. The asymmetric proximity method [43] is:

$$N(A, B) = 1 - \frac{1}{n} \sum_{i=1}^n |\mu_A^q(\mu_i) - \mu_B^q(\mu_i)|^{\frac{1}{t}} \quad (10)$$

In the formula,  $\mu_A(\mu_i)$  and  $\mu_B(\mu_i)$  are the membership degrees of objects corresponding to A and B, respectively, which belong to  $\mu_i$ . Among them,  $q$  plays a regulatory role in the calculation results and compensates with the role of  $1/t$ , which can help make the calculation results more conducive to classification.  $q > 0$  can be used, and the value should not be too large as this is not conducive to grading, and  $q=3$  is taken in this application study. If  $N(A, B) = \max N(A, B_i) (i=1, 2, \dots, n)$ , the emergency management capability level is  $e$ .

## 2.2.2. Index Weight Determination Method

### Traditional Principal Component Analysis

Principal component analysis is a mathematical method to reduce the dimension of a variety of sample data through certain mathematical means to improve the concentration of sample information [44]. The specific steps of principal component analysis are as follows:

- ① Calculate the covariance matrix. The covariance of normalized sample data is

$$s_{kj} = \frac{1}{n} \sum_{i=1}^n (X_{ik} - \bar{X}_k)(X_{ij} - \bar{X}_j) \quad (k, j = 1, 2, \dots, p) \quad (11)$$

In the formula,  $s_{kj}$  is the covariance value of the  $k$  evaluation index and the  $j$  evaluation index; and  $\bar{X}_k$  is the normalized sample mean of the  $k$  evaluation index.

- ② Calculate the eigenvalues and unit eigenvectors of the covariance matrix. Under the condition of data sample normalization, the covariance matrix is the correlation coefficient matrix. The eigenvalues  $\lambda_j$  ( $j = 1, 2, \dots, p$ ) and eigenvectors of the  $j$ th evaluation index of the correlation coefficient matrix are obtained using the Jacobian determinant method. The eigenvalues are sorted from large to small ( $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ ). The variance contribution rate is the proportion of a certain eigenvalue to the total number of eigenvalues:

$$G(\lambda_j) = \frac{\lambda_j}{\sum_{k=1}^p \lambda_k} \quad (j, k = 1, 2, \dots, p) \quad (12)$$

$G(\lambda_j)$  is the variance contribution rate of  $\lambda_j$ .

- ③ Select the principal component. The cumulative variance contribution rate of the principal component is

$$G(m) = \sum_{j=1}^m \lambda_j / \sum_{k=1}^p \lambda_k \quad (m, j, k = 1, 2, \dots, p) \quad (13)$$

In the formula,  $G(m)$  is the cumulative variance contribution rate of the first  $m$  eigenvalues. When  $G(m) > 85\%$ ,  $m$  is called the principal component.

- ④ Calculate principal component load:

$$l_{cj} = \sqrt{\lambda_c} a_{cj} \quad (c = 1, 2, \dots, m; j = 1, 2, \dots, p) \quad (14)$$

In the formula,  $F_c$  is the  $c$  principal component score, and  $X_p$  is the  $p$  evaluation index of the normalized sample matrix.

- ⑤ Calculation of principal component scores:

$$F_c = a_{c1}X_1 + a_{c2}X_2 + \dots + a_{cp}X_p \quad (c = 1, 2, \dots, m) \quad (15)$$

where  $F_c$  is the  $c$ th principal component score, and  $X_p$  is the  $p$ th evaluation indicator of the normalized sample matrix.

- ⑥ Calculate the composite principal component score:

$$F = \sum_{c=1}^m \frac{\lambda_c}{\sum_{d=1}^m \lambda_d} \times F_c \quad (c = 1, 2, \dots, m) \quad (16)$$

$F$  is the comprehensive principal component score of traditional principal component analysis.

### The Improvement of Principal Component Analysis using Entropy Weight Method

The entropy weight method is an objective weighting method to describe the irreversible phenomenon of molecules. The greater the difference between the parameters of a system, the more information it contains and the smaller the entropy value; in contrast, the greater the entropy value, to evaluate the contribution of a weight index to the system. The entropy weight method has the advantages of highlighting the local information of the system and being less affected by subjective factors and has been widely used in many engineering fields [45]. When the first principal component of the traditional principal component analysis method does not meet the requirement that the cumulative variance contribution rate is greater than 85%, multiple principal components need to be fused, and the weight distribution among the principal components is the main factor affecting the lithology stratification effect. The traditional principal component analysis method is to calculate the variance contribution rate of each principal component by weighting the principal components, but the principal components are independent of each other, and the information content of the calculation results may not rise but fall. In this paper, the entropy weight method is used to improve the traditional principal component analysis method. According to the variation degree of each principal component of the traditional principal component analysis method, the entropy weight method is used to recalculate the weight of each principal component. Finally, the comprehensive value of the entropy weight principal component is used as the index parameter of the haze risk division. The steps are as follows:

① Normalization calculation of each principal component:

$$F_{ic}^* = \frac{F_{ic} - F_{c(\min)}}{F_{c(\max)} - F_{c(\min)}} \quad (i = 1, 2, \dots, n; c = 1, 2, \dots, m) \quad (17)$$

In the formula,  $F_{ic}^*$  is the normalized score of the  $c$  principal component of the first sample,  $F_{ic}$  is the score of the  $c$  principal component of the  $i$  sample,  $F_{c(\max)}$  is the maximum score of the  $c$  principal component, and  $F_{c(\min)}$  is the minimum score of the  $c$  principal component.

② Calculate the proportion of each principal component sample:

$$E_c = -\frac{1}{\ln n} \sum_{i=1}^n p_{ic} \times \ln p_{ic} \quad (i = 1, 2, \dots, n; c = 1, 2, \dots, m) \quad (18)$$

$E_c$  is the entropy of the  $c$  principal component.

③ Calculate the weight of each principal component:

$$\omega_c = (1 - E_c) / \sum_{f=1}^m (1 - E_f) \quad (c, f = 1, 2, \dots, m) \quad (19)$$

where  $\omega_c$  is the weight of the  $c$  th principal component.

④ Calculate the entropy principal component composite score:

$$F' = \sum_{c=1}^m \omega_c \times p_{ic} \quad (i = 1, 2, \dots, n; c = 1, 2, \dots, m) \quad (20)$$

### 2.3. Calculate the Weight of Each Evaluation Index

As shown in Table 2, the KMO sampling fitness number for this principal component analysis is 0.684, which is greater than its threshold value of 0.5, indicating that there is correlation between the variables, which meets the requirements. The Sig value is 0.000, which is less than 0.05, which indicates that this data can be subjected to principal component analysis and is scientific and informative. The eigenvalues and contribution rates of each principal component are shown in Table 3. According to the principle that the eigenvalue is greater than 1, the first three items are selected as the main components, and



the variance contribution of these three items are 39.76, 22.74 and 15.39%, respectively, and the cumulative contribution of the three items is 77.9%, which can basically reflect the information of the original indexes. The first 3 items are used as principal component factors and denoted by F1, F2 and F3, so that the original 13 indicators are replaced by the first 3 principal components, and the loading status of each factor on the original indicators can be calculated at the same time.

**Table 2.** KMO test and Bartlett's test.

Bartlett's Sphericity Test	
KMO	0.684
Approximate cardinality	720.527
Df	66.000
P	0.000 ***

Note: \*\*\*, \*\* and \* represent 1, 5 and 10% significance levels, respectively.

**Table 3.** Principal component loading matrix.

Ingredients	Characteristic Root		
	Characteristic Value	Percentage of Variance	Accumulation
1	5.169	39.76%	39.76%
2	2.957	22.74%	62.50%
3	2.001	15.39%	77.90%
4	0.886	6.82%	84.71%
5	0.638	4.91%	89.62%
6	0.431	3.31%	92.94%
7	0.326	2.51%	95.44%
8	0.242	1.86%	97.30%
9	0.167	1.28%	98.59%
10	0.111	0.85%	99.44%
11	0.044	0.34%	99.78%
12	0.023	0.18%	99.96%
13	0.006	0.04%	100.00%

As seen in Table 3, the F1 eigenvalue is 5.169, with a contribution rate of 39.76%. As seen in Table 4, tops the three principal components and is the primary driver of haze risk formation in Fenwei Plain cities. Analysis of the principal component F1 loadings reveals that the 1st principal component F1 has large values above 0.66 for indicators X1 (economic density), X4 (share of coal consumption), X11 (population density) and X12 (number of health institutions), which indicates that economic development density, coal consumption, population density and health institutions are the first constituents of haze risk. The top 3 loadings of principal component F2 are X3 (number of motor vehicles), X5 (housing construction areas), X6 (annual average concentration of VOCs) and X8 (annual average concentration of SO<sub>2</sub>), which shows that the haze hazard in Fenwei Plain cities is mainly dominated by the toxic gases VOCs and SO<sub>2</sub>. The indicator with the highest principal component F3 loading value is X7 (annual average PM<sub>10</sub> concentration), followed by PM<sub>2.5</sub> and the share of secondary industry, reflecting that PM<sub>2.5</sub>, PM<sub>10</sub> and secondary industry are also important environmental factors in the formation of haze risk.

**Table 4.** Principal component loading matrix.

Indicators		F1	F2	F3
X1	Economic density	0.872	0.227	−0.339
X2	Percentage of secondary	−0.327	−0.445	0.482

	sector of economy			
X3	Number of motor vehicles	−0.476	0.755	−0.23
X4	Share of coal consumption	0.662	−0.552	0.216
X5	Area of housing construction	−0.541	0.674	−0.04
X6	Annual average VOCs concentration	0.625	0.681	−0.065
X7	Annual average PM <sub>10</sub> concentration	0.473	0.25	0.759
X8	Annual average SO <sub>2</sub> concentration	−0.414	0.63	0.425
X9	Annual average NO <sub>2</sub> concentration	0.644	0.119	0.365
X10	Annual average PM <sub>2.5</sub> concentration	0.246	0.511	0.666
X11	Population density	0.929	0.21	−0.153
X12	Number of health establishments	0.773	−0.218	0.141
X13	Area of greenery coverage	0.788	0.297	−0.416

#### 2.4. Haze Hazard Risk Principal Component Composite Score and Ranking

The loadings in the principal component loadings matrix reflect the extent to which the indicators play a role in the formation of haze risk, so the indicator weights can be expressed in terms of the indicator loadings. Using the weighted model, the scores of the evaluation units on the 3 principal components and the haze risk indices on the different principal components can be calculated as:

$$F_{iy} = \sum W_j X_j^i \quad (21)$$

where  $F_{iy}$  is the haze risk index of the  $i$ th evaluation unit on different principal components, e.g., in the 1st principal component, it is the haze risk index of evaluation unit  $i$  on F1.  $W_j$  is the loading value of the  $j$ th indicator on the corresponding principal component, and  $X_j^i$  is the standardized value of the  $j$ th indicator of the  $i$ th evaluation unit.

According to the principle of principal component analysis, the proportion of each principal component to the cumulative contribution reflects the importance of each principal component. The weight can be determined by analyzing the contribution of the principal components, and the weighting model can be used to calculate the comprehensive score of the evaluation unit, which is the comprehensive haze risk index  $F$ . It is calculated as:

$$F = \frac{(F_1 P_1 + F_2 P_2 + F_3 P_3)}{P_1 + P_2 + P_3} \quad (22)$$

where  $F_1$ ,  $F_2$  and  $F_3$  represent the scores of the first, second and third principal components, respectively, whereas  $P_1$ ,  $P_2$  and  $P_3$  represent their corresponding contribution rates.

#### 2.5. LightGBM

The light gradient lifter is a decision tree algorithm proposed based on gradient one-sided sampling and unique feature bundling with optimization in the negative gradient direction of the loss function [46]. The LightGBM algorithm is more efficient in processing high-dimensional big data due to the unique feature bundling (EFB) algorithm and gradient-based one-sided sampling (GOSS) algorithm in LightGBM. Suppose a training set  $Q$ ,  $Q = \{(x_i, y_i)\}_{i=1}^N$ , consisting of  $N$  samples, where  $x_i \in X = \{x_1, x_2, \dots, x_k\}$  represents the data,  $X$  denotes the  $k$ -dimensional vector space,  $y_i \in Y = \{0, 1\}$  represents the category labels, and  $y_i = 1$  denotes the faulty samples. The objective of the LightGBM algorithm is to find a mapping relation  $\bar{G}(x)$  to approximate the function  $G(x)$ , such that the loss function  $\phi(y, G(x))$  is minimized. The objective function can be expressed as:

$$L_t = \sum_{i=1}^m L(y_i, f_{t-1}(x_i) + h_t(x_i)) + \gamma J + \sum_k \Omega(f_k) \quad (23)$$

where  $L(y_i, f_{t-1}(x_i) + h_t(x_i))$  is the loss function, and  $\Omega(f_k)$  denotes the regular term unlike the fast descent method of GBDT.

LightGBM uses Newton method to quickly approximate the objective function. Equation can be derived as:

$$L_t \cong \sum_{i=1}^m (g_i f_t(x_i) + \frac{h_i f_t^2(x_i)}{2}) + \sum_k \Omega(f_k) \quad (24)$$

$g_i$  represents the first-order loss function, and  $h_i$  represents the second-order loss function. The equation is as follows:

$$g_i = \sigma_{F_{t-1}(x_i)} L(y_i, f_{t-1}(x_i)) \quad (25)$$

$$g_i = \sigma_{F_{t-1}(x_i)} L(y_i, f_{t-1}(x_i)) \quad (26)$$

The information gains in LightGBM are as follows:

$$H = \frac{1}{2} \left[ \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} + \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] \quad (27)$$

## 2.6. Improved PSO Algorithm

As a swarm intelligence algorithm, particle swarm optimization (PSO) has been widely used in various industries to solve practical problems in recent years. The traditional PSO algorithm is easily falls into local optimum and has poor convergence speed and accuracy in the iterative process. It is difficult to ensure the efficiency of the algorithm in practical engineering tasks. After using the topology, the optimization process of particle swarm is carried out as follows:

① All particles in the particle swarm are arranged from large to small according to the fitness value of the particles at the initial time. The first  $N$  particles are selected as the main particles,  $N$  is a positive integer, and all particles in the particle swarm except the main particles are used as the slave particles.

② The K-means clustering method is used to classify the subordinate particles by using each main particle as the clustering center.

③ Each master particle from the particle group and its corresponding cluster center is used as an improved particle group to obtain  $N$  improved particle groups.

④ Using Formulas (28) and (29) to update the velocity of the main particle and the slave particle in the improved particle group and multiple slave particle groups, each slave particle group has the same number of slave particles.

$$v_{mi(t+1)} = k \left[ v_{mi(t)} + \phi_1 \gamma_1 (mp_{i(t)}^{best} - x_{mi(t)}) + \phi_2 \gamma_2 (mg_{i(t)}^{best} - x_{mi(t)}) \right] \quad (28)$$

$$v_{sij(t+1)} = k \left[ v_{sij(t)} + \phi_1 \gamma_1 (sp_{ij(t)}^{best} - x_{sij(t)}) + \phi_2 \gamma_2 (sg_{i(t)}^{best} - x_{sij(t)}) \right] \quad (29)$$

In the formula,  $x_{mi(t)}$  represents the position record of the main particle in the  $i$  th improved particle group at the current time  $t$ .  $x_{sij(t)}$  denotes the position record of the  $j$  th slave particle in the  $i$  th improved particle group at the current time  $t$ .  $v_{mi(t)}$  denotes the velocity of the main particle in the  $i$  th improved particle group at the current time  $t$ .  $v_{sij(t)}$  represents the velocity value of the  $j$  th slave particle in the  $i$  th improved particle group

at the current time  $t$ .  $k$  is the convergence factor and a constant.  $\varphi_1$  and  $\varphi_2$  are learning factors.  $mp_{i(t)}^{best}$  represents the historical optimal position record of the main particle in the  $i$ th improved particle group at time  $t$  (taking the optimal position value of the main particle in the  $i$ th improved particle group).  $sp_{ij(t)}^{best}$  represents the historical optimal position record of the  $j$ th slave particle in the  $i$ th improved particle group at time  $t$ .  $\gamma_1$  and  $\gamma_2$  are constants greater than 0 and less than 1.  $mg_{(t)}^{best}$  represents the optimal historical position record of the primary particles in all improved particle groups at the current time  $t$ .  $sg_{i(t)}^{best}$  denotes the optimal record of the historical position of the particle in the  $i$ th improved particle group at the current time  $t$ .

⑤ The positions of master and slave particles are updated using Formulas (27) and (28) according to the updated velocity value:

$$x_{mi(t+1)} = x_{mi(t)} + v_{mi(t+1)} \quad (30)$$

$$x_{sij(t+1)} = x_{sij(t)} + v_{sij(t+1)} \quad (31)$$

⑥ According to the updated position parameters of the master-slave particles, the fitness value is recalculated, and iterative optimization is performed. Through the above process, it can be concluded that when the proposed improved particle swarm relationship topology is adopted, other main particle swarms can also jump out of the local extremum as much as possible to search for the global optimum and improve the accuracy of parameter optimization when the slave particle group in a certain region falls into the local optimum. On the other hand, in the traditional particle swarm optimization algorithm,  $\varphi_1$  and  $\varphi_2$  are fixed values, generally taking a constant between 0 and 2, which limits the global and local search ability of particles to a certain extent. Therefore, linear increasing and decreasing functions are introduced to improve this part. The improved formula is as follows:

$$\varphi_1(t) = \varphi_{01} \times \left(1 - \frac{t}{G}\right) \quad (32)$$

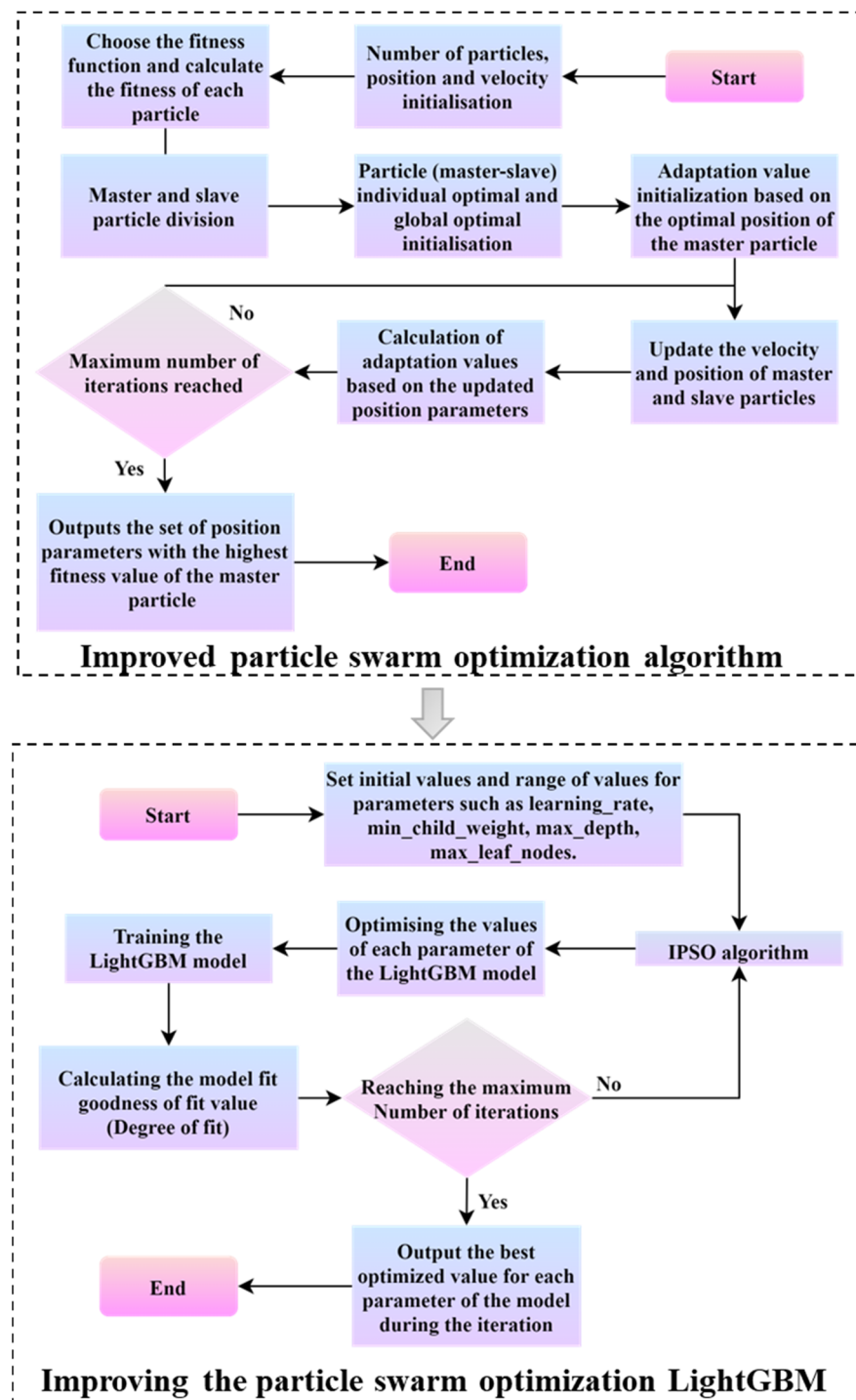
$$\varphi_2(t) = \varphi_{02} \times \frac{t}{G} \quad (33)$$

In the formula,  $\varphi_{01}$  and  $\varphi_{02}$  are the initial values of the learning factor.  $t$  and  $G$  are the current and maximum number of iterations, respectively. The improved master-slave particle velocity update formula is as follows:

$$v_{mi(t+1)} = k \left[ v_{mi(t)} + \varphi_1(t) \gamma_1 (mp_{i(t)}^{best} - x_{mi(t)}) + \varphi_2(t) \gamma_2 (mg_{(t)}^{best} - x_{mi(t)}) \right] \quad (34)$$

$$v_{sij(t+1)} = k \left[ v_{sij(t)} + \varphi_1(t) \gamma_1 (sp_{ij(t)}^{best} - x_{sij(t)}) + \varphi_2(t) \gamma_2 (sg_{i(t)}^{best} - x_{sij(t)}) \right] \quad (35)$$

In summary, the IPSO-LightGBM model construction process is shown in Figure 2.



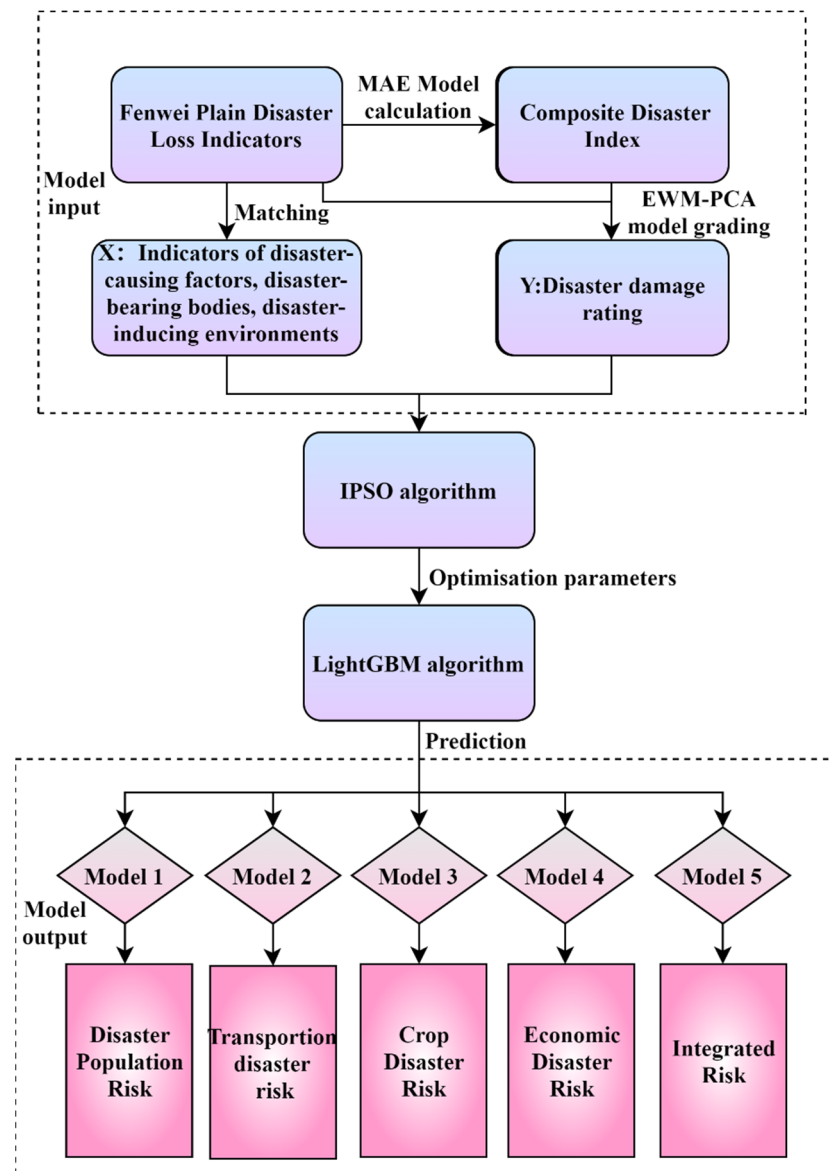
**Figure 2.** IPSO-LightGBM model construction process.

### 2.7. Model Establishment and Performance Evaluation Index

Based on the IPSO-LightGBM algorithm, this paper establishes five risk assessment models, including population, transportation, crop and economic disaster risks and integrated risk. All models take the three types of indicators of disaster-causing factors, disaster-pregnant environments and disaster-affected bodies as input vectors, and different loss risk levels as output vectors. The specific model establishment process is shown in Figure 3.

The LightGBM model optimized using IPSO has higher fitness values during the iterative process and converges faster than the PSO algorithm. In the iterative search

process, the search stability of IPSO algorithm is high, and the optimal hyperparameter combination of LightGBM model has been searched for around 300 iterations. In contrast, the PSO algorithm is more volatile in the iterative search process, and the convergence does not appear at 600 iterations, and the gap between the adaptation degree and IPSO algorithm is further widened, and the final convergence adaptation degree of the experiment is lower than that of the IPSO algorithm. Under the given termination iteration condition, the model parameters obtained using IPSO optimization are Learning\_rate = 0.25, gamma = 0.13, max\_depth = 7, min\_child\_weight = 3, and lambda = 1.



**Figure 3.** Haze disaster risk assessment model building process.

Among them, the model tuning parameters were optimized using 10-fold cross-test, and the grid search was performed for the main three parameters of the XGBoost model, which are the number of weak classifiers, the maximum depth of the decision tree and the learning rate. The six model optimal parameters and the accuracy of the training set are shown in Table 5.

**Table 5.** Assessment results of different risk levels of the validation set.

Evaluation Model	Number of Weak Classifiers	Maximum Depth	Learning Rate	Training Set Accuracy
Disaster population risk	900	4	0.1	0.977
Transportation disaster risk	800	2	0.2	0.95
Crop disaster risk	1000	3	0.1	0.987
Economic disaster risk	900	4	0.2	0.948
Integrated risk	900	4	0.1	0.937

To evaluate the model accuracy, four evaluation indicators were selected, including accuracy (ACC), detection rate (P), recall (R) and F-value (F):

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (36)$$

$$Precision = \frac{T_P}{T_P + E_P} \quad (37)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (38)$$

$$F_{Measure} = \frac{2Precision \times Recall}{Precision + Recall} \quad (39)$$

where  $T_P$  means true positive, which is itself a positive sample, and the prediction is also a positive sample.  $T_N$  means true negative, which is itself a negative sample, and the prediction is also a negative sample.  $F_P$  means false positive, which is itself a positive sample, and the prediction is a negative sample.  $F_N$  denotes false negative, which is itself a negative sample, and the prediction is a positive sample. In the above evaluation index, the accuracy rate indicates the proportion of all correctly predicted samples to the total sample. Accuracy indicates the proportion of samples with positive predictions that are true positive samples. Recall indicates the proportion of positive cases in the actual sample that are correctly predicted. The F-value is an indicator that balances the accuracy and recall rates and is the summed average of the two.

### 3. Results and Discussion

#### 3.1. Analysis of Evaluation Results

As seen from Table 6, F1, F2 and F3 represent the scores of the three principal component analyses, respectively, which are calculated by Equation (21). F is the composite haze risk index, which is calculated by Equation (22). The haze risk index is high in the Fenwei Plain urban agglomeration, especially in Xi'an, which is as high as 9.773. The analysis of the three principal component scores of the Fenwei Plain urban agglomeration reveals that the scores in F1 are much larger than those in F2 and F3, indicating that the main drivers are economic density, the number of motor vehicles, housing construction areas and the number of health institutions. Xi'an has been the center of economic development in Northwest China and is rich in industrial and mineral-rich resources, and urbanization is also rapid, resulting in increased environmental pressures, serious air pollution, the proximity of cities and the influence of the spatial spillover effects of pollutants, making it an extremely high-risk area for haze. To effectively control the risk of haze, the Fenwei Plain urban agglomeration should actively transform its economic development, improve traffic laws and regulations and regulate motor vehicles, thus reducing motor vehicle exhaust pollution and relieving the pressure on traffic. City governments should strengthen joint control and prevention mechanisms to control the construction area of an area within

a certain period of time through macro regulation so that it does not gather too much, make good isolation measures to reduce respirable particulates brought about by construction work, control the source of haze components and accelerate the improvement of urban greening. Urban greening can, to a certain extent, absorb harmful gases and dust generated by the city, reduce air pollution, purify the air and help reduce the vulnerability of disaster-bearing bodies. Xi'an, Sanmenxia, Lvliang and Luoyang belong to the high-risk area of haze. Analyzing the scores of their three principal components revealed that Xi'an, Sanmenxia and Luoyang are in the first principal component, indicating that the haze risk drivers in these two cities have the first principal component of economic density, the number of motor vehicles and housing construction areas. First, to carry out haze risk prevention and control, we must adjust the energy structure and promote clean energy. We must replace coal with other clean energy sources to reduce coal consumption, thus reducing the material components of haze formation and the risk of haze. Second, in the case of irreplaceable coal, we must improve the desulfurization of coal, denitrification and dust removal technology to reduce the emissions of sulfide and other emissions due to burning coal and achieve the purpose of reducing the risk of haze. Third, relevant enterprises should increase investment in research and development and actively develop new technologies to improve energy utilization, reduce energy consumption and achieve energy conservation, emission reduction and green development. The middle-risk areas are Lvliang, Luoyang, Linfen, Yuncheng and Baoji. Both Tongchuan and Jinzhong have high scores with the second principal component, while the SO<sub>2</sub> concentration, PM<sub>10</sub> concentration and coal consumption share of the second principal component also play an important role in the formation of haze risk. Weinan, Xianyang, Tongchuan and Jinzhong are four cities with relatively high scores for the third principal component; with rich forest vegetation, the strong self-cleaning ability of the atmosphere and a high rate of good air, their haze pollution is small and low risk. Therefore, small enterprises with low capacity and high emissions should be eliminated by strengthening the regulation of pollutant emissions from factories. The approval system of enterprise project engineering should be established, improved and strictly enforced, raising the threshold of enterprise access and controlling the emission of haze material components from the source.

**Table 6.** Results of haze disaster risk evaluation in Fenwei Plain cities.

	F1	F2	F3	F	Rank
Xi'an	9.49	1.161	0.284	9.773159	1
Sanmenxia	8.94	1.585	0.272	9.371132	2
Lvliang	8.585	1.927	0.38	9.199759	3
Luoyang	8.946	1.087	0.593	9.189809	4
Linfen	8.384	1.65	0.716	8.874841	5
Yuncheng	8.354	1.637	0.668	8.839851	6
Baoji	8.152	1.596	0.786	8.631265	7
Weinan	7.716	2.012	0.972	8.44993	8
Xianyang	6.998	2.296	1.442	8.389519	9
Tongchuan	6.658	2.751	1.143	8.359984	10
Jinzhong	6.059	3.044	0.781	7.914102	11

### 3.2. Case Verification

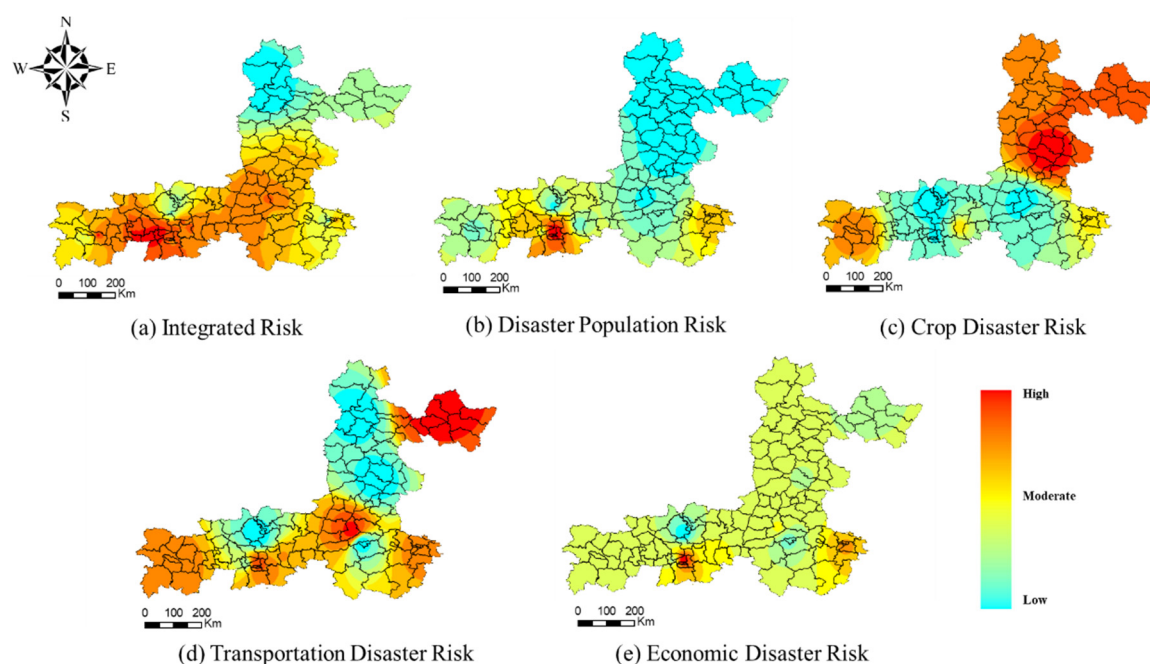
In this paper, we use the large-scale haze pollution in Fenwei Plain from November 1, 2021 to December 31, 2021 as a case study to validate the application of the major haze hazard assessment model based on the ISPO-LightGBM algorithm. This haze process caused a massive haze disaster in 11 prefecture-level cities in the Fenwei Plain, affecting a total of 52 million people. The established models were used to evaluate the affected population, transportation, crop and economic disaster risks and integrated risk, and were then compared with the actual disaster loss levels at the county level, and the results are



shown in Table 7. The five types of disaster risks in the Fenwei Plain are shown in Figure 4.

**Table 7.** Assessment results of different risk levels of the validation set.

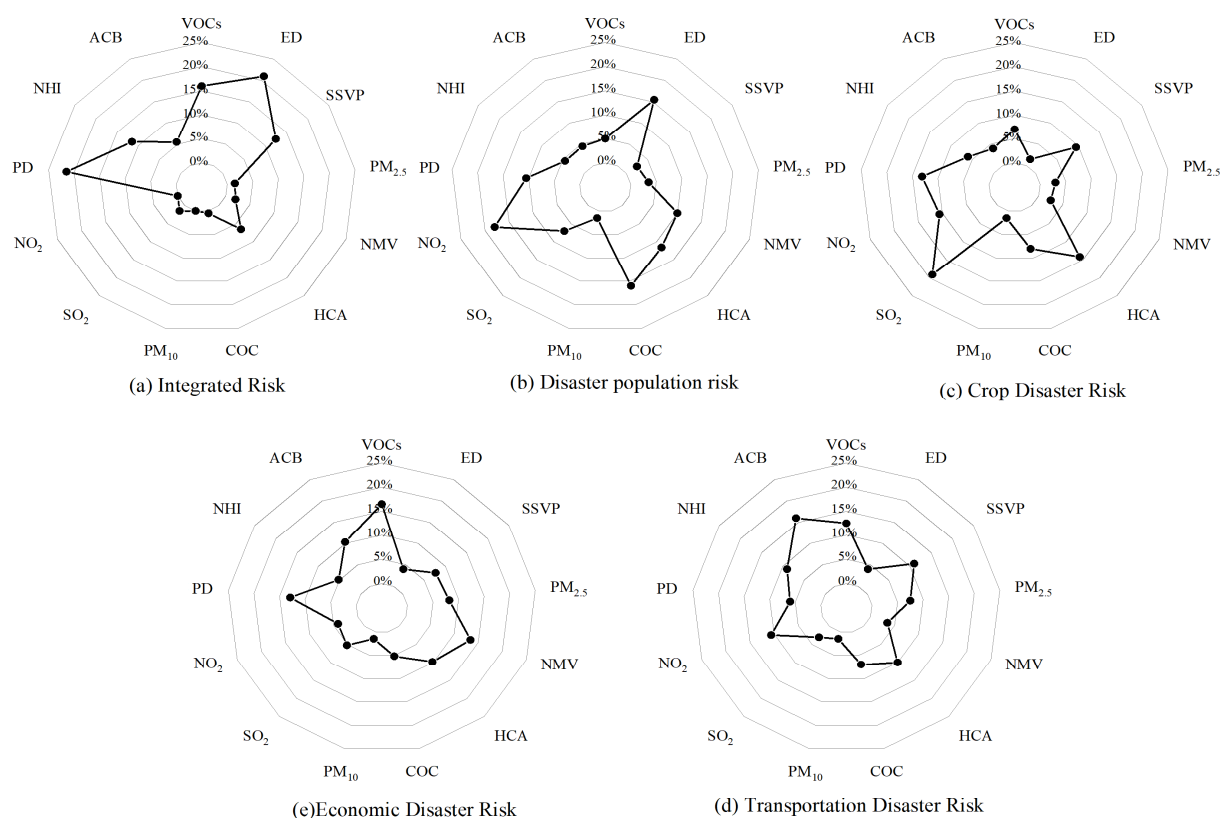
Evaluation Model	Accuracy	Precision	Recall	F
Disaster population risk	0.85	0.88	0.82	0.85
Transportation disaster risk	0.95	0.94	0.93	0.92
Crop disaster risk	0.84	0.89	0.85	0.87
Economic disaster risk	0.65	0.64	0.65	0.63
Integrated risk	0.86	0.91	0.93	0.92



**Figure 4.** Five disaster risks in Fenwei Plain.

### 3.3. Importance of Indicators

To understand the various factors that influence the assessment results, it is necessary to calculate the specific contribution of each assessment indicator. The LightGBM algorithm calculates the importance of an indicator based on the principle that the more times an input indicator is selected as a branching feature when the decision tree branches, the more important the feature is. In this paper, the importance of indicators was calculated for each of the five types of risk assessment objectives, and the results are shown in Figure 5.



**Figure 5.** Importance of indicators for different risk assessment types.

Among them, the most important indicators affecting the risk assessment of the affected population are  $\text{NO}_2$ , economic density and coal consumption, indicating that the risk of disastrous factors, environmental sensitivity of disaster and disaster-bearing body all contribute to the assessment results. The transportation disaster risk has a greater relationship with  $\text{NO}_2$  and VOCs in green areas in built-up areas, indicating that transportation risk has a greater correlation and impact with pregnant environments. The crop disaster risk has a great relationship with  $\text{SO}_2$ , the proportion of secondary industry, housing construction areas and population density; in particular, the impact of  $\text{SO}_2$  is prominent, indicating that the disaster of crops is closely related to the disaster environment. The main influencing factors of economic disaster risk are VOCs, population density and the number of motor vehicles. The economic disaster risk is closely related to disaster-causing factors and disaster-pregnant environments. The main influencing factors of integrated risk are population and GDP densities. The possible reason is that GDP itself is a comprehensive index. GDP cannot only reflect the comprehensive exposure of the disaster-bearing body in the region but also the vulnerability of the disaster-bearing body in the region to a certain extent. In other words, it can be considered that the comprehensive disaster prevention and mitigation capacity of a region with high GDP is stronger than that of the region with low GDP. Overall, the contribution of different indicators to different risk assessment results is not the same, and none of the indicators can contribute to a negligible extent, with the contribution of each indicator ranging from 5 to 12%.

### 3.4. Impact of Indicator Size on Assessment Results

As shown in Table 8, RD denotes the risk of disastrous factors, ESD denotes the environmental sensitivity of disaster, and DBB denotes the disaster-bearing body. To examine the influence of the number of indicators on the accuracy of the assessment model, this paper combined the input indicators of different dimensions and compared the accuracy in the haze risk assessment results using only the causative factor, the combination of the risk of disastrous factors, the environmental sensitivity of disaster, the disaster-bearing body and the use of all indicators. By comparison, it was found that the change in the number of indicators had less impact on the assessment results of the two models of population and transportation disaster risks. However, the number of indicators has a large impact on the accuracy of the assessment of the three models of crop and economic disaster risks and integrated risk, and the accuracy is the lowest if the model input is only the causative factor, which is 4–16% lower than the full indicator. In addition to the disaster-causing factors, the addition of both the environmental sensitivity of disaster and disaster-bearing body indicators will improve the accuracy, and the disaster-bearing body indicators will improve more than the environmental sensitivity of the disaster indicators because the disaster-bearing body indicators have more subcomponents. The highest accuracy rate was achieved by using all indicators together as input, indicating that the amount of indicators has a significant impact on the assessment results.

**Table 8.** Accuracy of risk assessment for different indicator quantities.

Index	RD	RD + ESS	RD + DBB	RD + EES + DBB
Disaster population risk assessment accuracy/%	79	81	81	85
Transportation disaster risk assessment accuracy/%	91	92	94	95
Crop disaster risk assessment accuracy/%	73	78	80	84
Economic disaster risk assessment accuracy/%	48	52	57	65
Integrated risk assessment accuracy/%	72	75	80	86

### 3.5. Impact of Environmental Factor Variables on Assessment Results

We added the meteorological conditions as well as the terrain as a factor for comparison. Pearson analysis was performed for topographic and meteorological factors, as shown in Table 9. Altitude, temperature and wind speed were found to be moderately positively correlated with haze risk. Other factors such as woodland, grassland, relative humidity and precipitation showed weak negative correlations.

As shown in Table 10, we added topographic and meteorological factors to each of the five integrated models, and we found that the topographic factor enhances the model less, and the meteorological factor enhances the model significantly relative to the topographic factor. The combined input of topographic and meteorological factors improves the models more significantly. Compared with the model before input, the accuracy of the five risk models was improved by 6.12% on average.

**Table 9.** Analysis and description of environmental factor variables.

Environmental Factors	Variable Name	Unit	Variable Description	Pearson Correlation
Topographical factors (TF)	Altitude	M	Monitoring station altitude	−0.559
	for_X	%	Forest	−0.379
	gra_X	%	Grass	−0.299
	SSD	H	Sunshine hours	0.018
	WD	-	Wind direction	0.202
Meteorological factors (MF)	TEM	°C	Temperatures	0.523
	RHU	%	Relative humidity	−0.215
	PRE	mm	Precipitation	−0.346
	WIN	m·s <sup>−1</sup>	Wind speed	0.415

**Table 10.** Accuracy of risk assessment for different indicator quantities.

Index	Topographical Factors	Meteorological Factors	Comprehensive Factors (TF + MF)
Disaster population risk assessment accuracy/%	86	88	90
Transportation disaster risk assessment accuracy/%	95.7	95	96
Crop disaster risk assessment accuracy/%	84.6	86	88
Economic disaster risk assessment accuracy/%	67	69	72
Integrated risk assessment accuracy/%	88	89	93

### 3.6. Comparison with other Studies

A comparison of the risk assessment model proposed in this study with models proposed in other similar studies can better elucidate the differences between this study and other studies, as shown in Table 11. Currently, there are few detailed studies on haze risk assessment. Second, in terms of feature selection methods, this paper uses the new gradient enhancement algorithm LightGBM to filter the features. To the best of our knowledge, there are few studies using the LightGBM algorithm to filter features. Compared with other mainstream integration algorithms in the boosting family, optimizing the LightGBM model using ISPO requires less parameter tuning, shows faster adaptation to the model and is more scalable. In conclusion, compared with the models proposed in other studies, the model proposed in this paper can effectively solve the haze risk assessment problem and has good prediction performance, especially with a precision of 0.91.

**Table 11.** Comparison with other studies.

Algorithm	City	Accuracy	Precision	Recall	F
PSO-SVM [47]	Beijing	0.82	0.91	0.75	0.82
Efficient weighted naive bayes classifiers [48]	Delhi	0.836	0.87	0.82	0.873
MCS-RF [49]	Beijing	0.828	0.875	0.872	0.871
APNet [50]	Beijing-Tianjin-Hebei	0.848	0.846	0.789	0.817
CNN [51]	Taipei	-	0.87	0.84	-
PCA-MEE-ISPO-LightGBM	Fenwei Plain	0.86	0.91	0.93	0.92

## 4. Conclusions

In this paper, based on nearly 300,000 indicators of haze cases in 11 cities in Fenwei Plain in China, a haze disaster assessment model is established using the PCA-MEE-ISPO-LightGBM algorithm, and the model is validated with data from the haze pollution process in Fenwei Plain region in mid-November 2021. The results show that the model can be used for the assessment of the affected population, transportation, crop and economic disaster risks and integrated risk before major haze disaster events, which is important for disaster risk management operations.

(1) Through the matter-element analysis, we construct the classical domain, determine the matter-element to be evaluated and calculate the correlation degree of the evaluation index and the haze disaster assessment level. Introducing the asymmetric closeness degree criterion, the index weight is improved using the entropy weight method to the principal component analysis method, and the haze disaster evaluation method based on the matter-element extension model of the improved principal component analysis is proposed. The IPSO optimization algorithm which divides the topological relationship between master and slave particles and dynamically adjusts the iterative learning factor is proposed to solve the problem that the particle swarm easily falls into the local optimal region in the iterative process. The IPSO is integrated into the parameter optimization process of the LightGBM model, and the hyperparameters of the LightGBM prediction model are optimized. The disaster risk assessment models based on the PCA-MEE-ISPO-LightGBM algorithm show good applicability. The performance indexes of the five

models in the risk assessment, such as accuracy, detection rates, recall rates and F-values, are above 80%, indicating that the models show good generalization performance and can be used in actual disaster risk assessment work.

(2) The average annual concentration of VOCs, economic density, number of motor vehicles, housing construction areas, average annual concentration of SO<sub>2</sub> and PM<sub>2.5</sub>, and the share of secondary industry and coal consumption have a strong influence on the risk of urban haze disaster. The higher the level of economic development in a city, the more motor vehicles, the higher the dependence of economic development on the secondary industry, the higher the coal consumption and the higher the risk of urban haze. Xi'an has the highest risk of haze disaster, and Jinzhong has the lowest risk of haze. The haze hazard risk degree of Fenwei Plain has obvious geographical differences, and the haze risk of Xi'an urban agglomeration is extremely high and centers on it, gradually decreasing roughly in the west, south and north directions.

(3) The model can calculate the contribution of importance evaluation indicators to the risk assessment results. In addition to the influence of VOC indicators on most of the assessment targets, different risk assessment targets have different influencing factors. Economic disaster risk is influenced by the factors of the disaster-bearing body. The affected population, crop and economic disaster risks are mainly influenced by the environmental sensitivity of disaster, whereas the main influencing factors of integrated disaster risk are population and GDP densities. The importance of indicators increases the interpretability of the risk assessment models, improves the understanding of the relationship between indicators and assessment results, and helps improve the understanding of the "black box" model of machine-learning algorithms.

(4) The amount of indicators and sample size play an important role for data-driven assessment models. Integrated learning algorithms in disaster risk assessment downplay hazard mechanisms such as hazards and vulnerabilities and purely use disaster system-related data for learning and simpler modeling, which also requires the sufficient accumulation of assessment indicators and sample size. On the one hand, hazard-causing factor indicators, hazard-inducing environment indicators and hazard-bearing body indicators all have an important impact on the results of hazard risk assessment, and the use of the full indicator volume can improve the accuracy of assessment by 10–15% compared with only the hazard-causing factor indicators. On the other hand, increasing the sample size by one to two orders of magnitude can improve the assessment accuracy by 5–13%. This indicates that disaster big data can be of great help to improve the performance of disaster risk assessment models.

A disaster risk assessment model for the haze process in the Fenwei Plain is established using disaster big data. With rapid socio-economic development, the regional disaster-bearing body and environmental sensitivity of disaster will undergo many changes. In future research, it is necessary to continuously introduce the latest data, update and accumulate big data, and improve the reliability of the model. To summarize the next step, the focus is on two directions. The first is to continue to improve the indicator system and sample distribution, update the indicators using the first national comprehensive natural disaster risk census data and further improve the model. The second is to collect cases of major haze disaster processes Y.R.: editing, revision. in other regions and verify whether the model is generalizable in the Beijing-Tianjin-Hebei and Yangtze River Delta regions.

**Author Contributions:** H.D.: conceptualisation, methodology, modelling, writing original draft preparation. G.H.: revision. H.Z.: writing reviewing and editing. R.Y.: editing, revision. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was supported by the National Natural Science Foundation of China (71874134), Key technologies of human-computer intelligent interactive angle control terminal manufacturing and industrial cluster application (AA21077018-2), 2022 Guangxi Science and Technology Teacher's College Research Fund Key Research Project A(GXKS2022ZD001) collaborative control technology and path planning in automotive welding.

**Data Availability Statement:** Economic density and population density are from the Data Center for Resource and Environmental Sciences, Chinese Academy of Sciences, 2016–2021 (<https://www.resdc.cn/Default.aspx>) (accessed on 17 November 2022). The annual average concentrations of PM<sub>10</sub>, PM<sub>2.5</sub>, SO<sub>2</sub>, VOCs and NO<sub>2</sub> for 2016–2021 were calculated from the daily data downloaded from the National Real-Time Urban Air Quality Release Platform (<http://106.37.208.233:20035/>) (accessed on 17 November 2022).

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

## References

1. GBD 2017 Risk Factor Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2018**, *392*, 1923–1994. [https://doi.org/10.1016/S0140-6736\(23\)366-8](https://doi.org/10.1016/S0140-6736(23)366-8).
2. World Health Organization (WHO). Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease. 2016. Available online: [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health) (accessed on 16 November 2022).
3. Rohde, R.A.; Muller, R.A. Air pollution in China: Mapping of concentrations and sources. *PLoS ONE* **2015**, *10*, e0135749. <https://doi.org/10.1371/journal.pone.0135749>.
4. Zhang, X.; Wu, L.; Zhang, R.; Deng, S.; Zhang, Y.; Wu, J.; Li, Y.; Lin, L.; Li, L.; Wang, Y.; et al. Evaluating the relationships among economic growth, energy consumption, air emissions and air environmental protection investment in China. *Renew. Sustain. Energy Rev.* **2013**, *18*, 259–70. <https://doi.org/10.1016/j.rser.2012.10.029>.
5. Wang, S.; Hao, J. Air quality management in China: Issues, challenges, and options. *J. Environ. Sci.* **2012**, *24*, 2–13. <https://doi.org/10.1016/S1001-074260724-9>.
6. Guo, S.; Hu, M.; Zamora, M.L.; Peng, J.; Shang, D.; Zheng, J.; Du, Z.; Wu, Z.; Shao, M.; Zeng, L.; et al. Elucidating Severe Urban Haze Formation in China. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17373–17378. <https://doi.org/10.1073/pnas.1419604111>.
7. Huang, R.-J.; Zhang, Y.; Bozzetti, C.; Ho, K.-F.; Cao, J.-J.; Han, Y.; Daellenbach, K.R.; Slowik, J.G.; Platt, S.M.; Canonaco, F.; et al. High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **2014**, *514*, 218–222. <https://doi.org/10.1038/nature13774>.
8. Wang, P.; Wang, T.; Ying, Q. Regional source apportionment of summertime ozone and its precursors in the megacities of Beijing and Shanghai using a source-oriented chemical transport model. *Atmos. Environ.* **2020**, *224*, 117337. <https://doi.org/10.1016/j.atmosenv.2020.117337>.
9. Wu, W.; Zhao, B.; Wang, S.; Hao, J. Ozone and secondary organic aerosol formation potential from anthropogenic volatile organic compounds emissions in China. *J. Environ. Sci.* **2017**, *53*, 224–237.
10. Liu, Y.; Li, X.; Tang, G.; Wang, L.; Lv, B.; Guo, X.; Wang, Y. Secondary organic aerosols in Jinan, an urban site in North China: Significant anthropogenic contributions to heavy pollution. *J. Environ. Sci.* **2019**, *80*, 107–115. <https://doi.org/10.1016/j.jes.2018.11.009>.
11. Wu, R.; Li, J.; Hao, Y.; Li, Y.; Zeng, L.; Xie, S. Evolution process and sources of ambient volatile organic compounds during a severe haze event in Beijing, China. *Sci. Total Environ.* **2016**, *560–561*, 62–72. <https://doi.org/10.1016/j.scitotenv.2016.04.030>.
12. Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **2015**, *525*, 367–371. <https://doi.org/10.1038/nature15371>.
13. Lim, S.S.; Vos, T.; Flaxman, A.D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; AlMazroa, M.A.; Amann, M.; Anderson, H.R.; Andrews, K.G.; et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380*, 2224–2260. [https://doi.org/10.1016/S0140-6736\(12\)61766-8](https://doi.org/10.1016/S0140-6736(12)61766-8).
14. Silva, R.A.; West, J.J.; Zhang, Y.; Anenberg, S.C.; Lamarque, J.-F.; Shindell, D.; Collins, W.; Dalsøren, S.; Faluvegi, G.; Folberth, G.; et al. Global premature mortality due to anthropogenic outdoor air pollution and the contribution of past climate change. *Environ. Res. Lett.* **2013**, *8*, 034005. <https://doi.org/10.1088/1748-9326/8/3/034005>.
15. Zhang, L.; Li, X.; Chen, H.; Wu, Z.; Hu, M.; Yao, M. Haze Air Pollution Health Impacts of Breath-Borne VOCs. *Environ. Sci. Technol.* **2022**, *56*, 8541–8551. <https://doi.org/10.1021/acs.est.2c01778>.
16. Gao, J.; Woodward, A.; Vardoulakis, S.; Kovats, S.; Wilkinson, P.; Li, L.; Xu, L.; Li, J.; Yang, J.; Cao, L.; et al. Haze, public health and mitigation measures in China: A review of the current evidence for further policy response. *Sci. Total Environ.* **2017**, *578*, 148–157. <https://doi.org/10.1016/j.scitotenv.2016.10.231>.
17. Wang, X.; Tian, G.; Yang, D.; Zhang, W.; Lu, D.; Liu, Z. Responses of PM<sub>2.5</sub> pollution to urbanization in China. *Energy Policy* **2018**, *123*, 602e610. <https://doi.org/10.1016/j.enpol.2018.09.001>.
18. Chen, J.; Zhou, C.; Wang, S.; Li, S. Impacts of energy consumption structure, energy intensity, economic growth, urbanization on PM<sub>2.5</sub> concentrations in countries globally. *Appl. Energy* **2018**, *230*, 94e105. <https://doi.org/10.1016/j.apenergy.2018.08.089>.

19. Singh, N.; Murari, V.; Kumar, M.; Barman, S.C.; Banerjee, T. Fine particulates over South Asia: Review and meta-analysis of PM<sub>2.5</sub> source apportionment through receptor model. *Environ. Pollut.* **2017**, *223*, 121–136. <https://doi.org/10.1016/j.envpol.2016.12.071>.
20. Latif, M.T.; Othman, M.; Idris, N.; Juneng, L.; Abdullah, A.M.; Hamzah, W.P.; Khan, F.; Sulaiman, N.M.N.; Jewaratnam, J.; Aghamohammadi, N.; et al. Impact of regional haze towards air quality in Malaysia: A review. *Atmos. Environ.* **2018**, *177*, 28–44. <https://doi.org/10.1016/j.atmosenv.2018.01.002>.
21. Zhang, X.; Yu, B. Causality Analysis and Risk Assessment of Haze Disaster in Beijing. *Appl. Sci.* **2022**, *12*, 9291. <https://doi.org/10.3390/app12189291>.
22. Xu, B.; Luo, L.; Lin, B. A dynamic analysis of air pollution emissions in China: Evidence from nonparametric additive regression models. *Ecol. Indic.* **2016**, *63*, 346–358. <https://doi.org/10.1016/j.ecolind.2015.11.012>.
23. Ong, B.T.; Sugiura, K.; Zettsu, K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM<sub>2.5</sub>. *Neural Comput. Appl.* **2016**, *27*, 1553–1566. <https://doi.org/10.1007/s00521-015-1955-3>.
24. Xue, T.; Zheng, Y.; Tong, D.; Zheng, B.; Li, X.; Zhu, T.; Zhang, Q. Spatiotemporal continuous estimates of PM<sub>2.5</sub> concentrations in China, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. *Environ. Int.* **2019**, *123*, 345–357. <https://doi.org/10.1016/j.envint.2018.11.075>.
25. Zhou, Q.; Jiang, H.; Wang, J.; Zhou, J. A hybrid model for PM<sub>2.5</sub> forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Sci. Total. Environ.* **2014**, *496*, 264–274. <https://doi.org/10.1016/j.scitotenv.2014.07.051>.
26. Zhou, W.; Wu, X.; Ding, S.; Ji, X.; Pan, W. Predictions and mitigation strategies of PM<sub>2.5</sub> concentration in the Yangtze River Delta of China based on a novel nonlinear seasonal grey model. *Environ. Pollut.* **2021**, *276*, 116614. <https://doi.org/10.1016/j.envpol.2021.116614>.
27. Zhou, H.; Zhang, F.; Du, Z.; Liu, R. A theory-guided graph networks based PM<sub>2.5</sub> forecasting method. *Environ. Pollut.* **2021**, *21*, 118569. <https://doi.org/10.1016/j.envpol.2021.118569>.
28. Zhu, T.; Shang, J.; Zhao, D. The roles of heterogeneous chemical processes in the formation of an air pollution complex and gray haze. *Sci. China Chem.* **2011**, *54*, 145–153. <https://doi.org/10.1007/s11426-010-4181-y>.
29. Sivakumar, B.; Wallender, W.W.; Horwath, W.R.; Mitchell, J.P. Nonlinear deterministic analysis of air pollution dynamics in a rural and agricultural setting. *Adv. Complex. Syst.* **2007**, *10*, 581–597. <https://doi.org/10.1142/S0219525907001288>.
30. Ma, T.; Duan, F.; He, K.; Qin, Y.; Tong, D.; Geng, G.; Liu, X.; Li, H.; Yang, S.; Ye, S.; et al. Air pollution characteristics and their relationship with emissions and meteorology in the Yangtze River Delta region during 2014–2016. *J. Environ. Sci.* **2019**, *83*, 8–20. <https://doi.org/10.1016/j.jes.2019.02.031>.
31. Dhaka, S.K.; Dang, C.; Kumar, V.; Panwar, V.; Dimri, A.P.; Singh, N.; Patra, P.K.; Hayashida, S. PM<sub>2.5</sub> diminution and haze events over Delhi during the COVID-19 lockdown period: An interplay between the baseline pollution and meteorology. *Sci. Rep.* **2020**, *10*, 13442. <https://doi.org/10.1038/s41598-020-70179-8>.
32. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.-H.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500. <https://doi.org/10.1126/science.1227079>.
33. Sethi, J.K.; Mittal, M. A new feature selection method based on machine learning technique for air quality dataset. *J. Stat. Manag. Syst.* **2019**, *22*, 697–705. <https://doi.org/10.1080/09720510.2019.1609726>.
34. Bagoulla, C.; Guillotreau, P. Maritime transport in the French economy and its impact on air pollution: An input-output analysis. *Mar. Policy* **2020**, *116*, 103818. <https://doi.org/10.1016/j.marpol.2020.103818>.
35. Mosedale, T.J.; Stephenson, D.B.; Collins, M.; Mills, T.C. Granger causality of coupled climate processes: Ocean feedback on the North Atlantic Oscillation. *J. Clim.* **2019**, *19*, 1182–1194. <https://doi.org/10.1175/JCLI3653.1>.
36. Amann, M.; Klimont, Z.; Wagner, F. Regional and global emissions of air pollutants: Recent trends and future scenarios. *Annu. Rev. Environ. Resour.* **2013**, *38*, 31–55. <https://doi.org/10.1146/annurev-environ-052912-173303>.
37. An, Z.; Huang, R.-J.; Zhang, R.; Tie, X.; Li, G.; Cao, J.; Zhou, W.; Shi, Z.; Han, Y.; Gu, Z.; et al. Severe haze in northern China: A synergy of anthropogenic emissions and atmospheric processes. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8657–8666. <https://doi.org/10.1073/pnas.1900125116>.
38. Wang, B.; Liu, S.; Du, Q.; Yan, Y. Long term causality analyses of industrial pollutants and meteorological factors on PM<sub>2.5</sub> concentrations in Zhejiang Province. In Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, China, 20–22 July 2018; pp. 301–305. <https://doi.org/10.1109/ICISCE.2018.00070>.
39. Kuerban, M.; Waili, Y.; Fan, F.; Liu, Y.; Qin, W.; Dore, A.J.; Peng, J.; Xu, W.; Zhang, F. Spatio-temporal patterns of air pollution in China from 2015 to 2018 and implications for health risks. *Environ. Pollut.* **2020**, *258*, 113659. <https://doi.org/10.1016/j.envpol.2019.113659>.
40. Olmo, N.R.S.; Saldiva, P.H.D.N.; Braga, A.L.F.; Lin, C.A.; Santos, U.D.P.; Pereira, L.A.A. A review of low-level air pollution and adverse effects on human health: Implications for epidemiological studies and public policy. *Clinics* **2011**, *66*, 681–690. <https://doi.org/10.1590/S1807-59322011000400025>.
41. Shi, P. Theory on disaster science and disaster dynamics. *J. Nat. Disasters* **2002**, *2002*, 1–9. <https://doi.org/10.13577/j.jnd.2002.0301>.

42. Wang, Y.; Yang, J.; Zhou, M.; Zhang, D.; Song, F.; Dong, F.; Zhu, J.; Liu, L. Evaluating the sustainability of China's power generation industry based on a matter-element extension model. *Util. Policy* **2021**, *69*, 101166. <https://doi.org/10.1016/j.jup.2021.101166>.
43. Wang, H.; Cao, Y. Evaluation of surrounding rock stability based on improved matter-element extension model with variable weight. *China Saf. Sci. J.* **2013**, *23*, 23–29. <https://doi.org/10.16265/j.cnki.issn1003-3033.2013.08.016>.
44. Huang, D.; Jiang, F.; Li, K.; Tong, G.; Zhou, G. Scaled PCA: a New Approach to Dimension Reduction. *Manag. Sci.* **2022**, *68*, 1678–1695. <https://doi.org/10.1287/mnsc.2021.4020>.
45. Lei, B.; Fan, J. Adaptive granulation Renyi rough entropy image thresholding method with nested optimization. *Expert Syst. Appl.* **2022**, *203*, 117378. <https://doi.org/10.1016/j.eswa.2022.117378>.
46. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3147–3155.
47. Liu, W.; Guo, G.; Chen, F.; Chen, Y. Meteorological pattern analysis assisted daily PM<sub>2.5</sub> grades prediction using SVM optimized by PSO algorithm. *Atmos. Pollut. Res.* **2019**, *10*, 1482–1491. <https://doi.org/10.1016/j.apr.2019.04.005>.
48. Sethi, J.K.; Mittal, M. Efficient weighted naive bayes classifiers to predict air quality index. *Earth Sci. Inform.* **2022**, *15*, 541–552. <https://doi.org/10.1007/s12145-021-00755-7>.
49. Feng, C.; Tian, Y.; Gong, X.; Que, X.; Wang, W. MCS-RF: Mobile crowdsensing-based air quality estimation with random forest. *Int. J. Distrib. Sens. Netw.* **2018**, *14*, 1550147718804702. <https://doi.org/10.1177/1550147718804702>.
50. Zhu, J.; Deng, F.; Zhao, J.; Zheng, H. Attention-based parallel networks (APNet) for PM<sub>2.5</sub> spatiotemporal prediction. *Sci. Total Environ.* **2021**, *769*, 145082. <https://doi.org/10.1016/j.scitotenv.2021.145082>.
51. Wang, W.-C.V.; Lin, T.-H.; Liu, C.-H.; Su, C.-W.; Lung, S.-C.C. Fusion of Environmental Sensing on PM<sub>2.5</sub> and Deep Learning on Vehicle Detecting for Acquiring Roadside PM<sub>2.5</sub> Concentration Increments. *Sensors* **2020**, *20*, 4679. <https://doi.org/10.3390/s20174679>.