symmetry

MDPI

Article

The Gray-Box Based Modeling Approach Integrating Both Mechanism-Model and Data-Model: The Case of Atmospheric Contaminant Dispersion

Bin Chen [1], Yiduo Wang [1,*], Rongxiao Wang [1], Zhengqiu Zhu [1], Liang Ma [1], Xiaogang Qiu [1] and Weihui Dai [2]

[1] College of Systems Engineering, National University of Defense Technology, 109 Deya Road, Changsha 410073, China; nudtcb9372@gmail.com (B.C.); wangrongxiao12@nudt.edu.cn (R.W.); zhuzhengqiu12@nudt.edu.cn (Z.Z.); maliang09a@foxmail.com (L.M.); michael.qiu@139.com (X.Q.)

[2] School of Management, Fudan University, Shanghai 200433, China; whdai@fudan.edu.cn

* Correspondence: wangyiduo14@nudt.edu.cn; Tel.: +86-1327-207-0151

Received: 12 December 2019; Accepted: 31 January 2020; Published: 6 February 2020

Abstract: With the profound understanding of the world, modeling and simulation has been used to solve the problems of complex systems. Generally, mechanism-models are often used to model the engineering systems following the Newton laws, and this kind of modeling approach is called white-box modeling; however, when the internal structure and characteristics of some systems are hard to understand, the black-box modeling based on statistic and data-modeling is often used. For most complex real systems, a single modeling approach can hardly describe the target system accurately. In this paper, we firstly discuss and compare the white-box and black-box modeling approaches. Then, to mitigate the limitations of these two modeling methods in mechanism-partially-observed systems, the gray-box based modeling approach integrating both a mechanism model and data model is proposed. In order to explain the idea of gray-box based modeling, the atmosphere dispersion modeling is studied in practical cases from two symmetric aspects. Specifically, the framework of data assimilation is used to illustrate the modeling from white-box to gray-box, while the Gauss features based Support Vector Regression (SVR) models are used to illustrate the modeling from black-box to gray-box. To verify the feasibility of the gray-box modeling method, we conducted both simulation experiments and real dataset symmetry experiments. The experiment results show the enhanced performance of the gray-box based modeling approach. In the end, we expect that this gray-box based modeling approach will be an alternative modeling approach for different existing systems.

Keywords: mechanism model; data model; gray-box modeling; atmosphere dispersion modeling

1. Introduction

Modeling and simulation has become a most important means to understand and change the target world due to its significant contribution to system analysis [1,2]. Driven by the urgent need of the engineering and recent advances in information technology, modeling and simulation has made great progress in these years, and it has been successfully applied in various symmetric fields, including engineering (such as aerospace and manufacturing), society, economy, military, etc. In the modeling and simulation process, the accurate modeling underlies the successive simulation, analysis and prediction of the system. In other words, the accuracy of system prediction is subject to the model's accuracy. It is widely accepted that the choice of the modeling technique depends on the target object (itself) and available research conditions [2].

For most engineering systems (e.g., the electronic system), the mechanism and structure of them are often explicit [3]. Therefore, a white-box model can be established based on the prior knowledge of

the system mechanism. For instance, the Resistance-Inductance-Capacitance (RLC) circuit system can be modeled by basic circuit laws (e.g., the Kirchhoff's law) and described by differential equations. Generally, the system mechanism includes the structure and the causality of the system. That is to say, the relationship of the input, system structure, and the output can be obtained from this knowledge (e.g., an event or action is a direct consequence of another). This property makes the white-box modeling suitable for the structural and diagnostic analysis of system [2]. However, it is usually unpractical to describe a target system totally by the system mechanism because the running process of most systems in the real world is not clear enough (especially the non-engineering system) [4]. In practice, many systems (e.g., models in society and economy fields) have a hard time acquiring the exact prior knowledge [5]. Fortunately, data-modeling provides an alternative way bypassing the system mechanism. Due to the development of big data these years, many researchers use data-modeling methods in various fields (including science, engineering, economic, industries and others), to predict the future behavior of the target system [6,7]. However, the data-based black-box model also has some deficiencies. Firstly, compared with mechanism-based modeling, the black-box model established by the data-modeling technique only describes the correlation of the data (e.g., the case of "beer and diapers" in the market) [8]. This deficiency causes that the relationship between the input and the output in the black-box model are hardly interpretable. Secondly, the data-modeling approach builds a model thoroughly relying on the data. This characteristic means that it cannot cope with anomalies and changing circumstances of the system [9]. Finally, the simplex data-modeling approach has accuracy limitation when the raw data cannot provide complete information of the system possibly. In summary, both modeling approaches are suitable for solving system-specific modeling tasks (as mentioned before). However, each approach has its limitations when facing complex real systems, e.g., wildfire spread [10] and traffic simulation [11].

To mitigate the disadvantages of single white-box or black-box approach, a new gray-box modeling methodology integrating both mechanism-model and data-model is proposed, which aims to guide people to build corresponding models for well describing gray-box systems. Significantly, when facing different target systems, suitable mechanism knowledge and different statistical as well as machine learning techniques can be applied in the gray-box modeling process. To demonstrate how mechanism-model and data-model are integrated together in specific scenarios, this paper takes the modeling of the air contaminant dispersion (ACD) as an example [12]. As a typical gray-box system, we have only partial knowledge (e.g., the Gaussian dispersion model) about ACD and thus cannot build a fully precise predicting model [13,14]. Currently, there are some theoretical basis for describing the mechanism of atmospheric dispersion, such as the computational fluid dynamics (CFD) based on the advection–diffusion equations [15,16], the Lagrangian Stochastic (LS) model underlay by the random walk [17,18], and the Gaussian model based on the statistic theory [19,20]. However, conventional mechanism-based models are usually static in their running process, due to the fixed model parameters even if the real atmospheric environment is always changing with time dynamically. To deal with the ACD modeling in the dynamic meteorological environment, some investigators employed data assimilation to build the dynamic data driven atmospheric dispersion model, which combines the mechanism model with dynamic observations [21–23]. Zheng et al. [23] took the assumed leakage accident of Daya Bay nuclear power plant as the case. In the simulated case, the observation of several monitoring stations is assimilated into the Monte Carlo dispersion model by the ensemble Kalman filter (EnKF). The results illustrate that the ACD model with data assimilation outperforms the conventional ACD model in the prediction accuracy. Reddy et al. [22] built a dimension-variant atmospheric dispersion model by integrating the RIMPUFF model and particle filter. The experiment results show that the proposed model has more accurate prediction than both the conventional model and the extended Kalman filter based model. This research shows that the gray-box modeling approach which integrates data assimilation (black-box) with a mechanism (white-box) model is an effective way of the atmospheric dispersion in a dynamic meteorological environment. The improvement of data condition drives more researchers to pay more attention to the data-modeling approach. Among

the data-modeling approaches, machine learning is widely applied into the prediction of the air contaminant dispersion. Support Vector Machine (SVM) is a typical machine learning approach [24–26]. Ma et al. [27] compared and tested several machine learning approaches for ACD modeling in a well-known field dataset Project Prairie Grass, including the SVM approach. Yeganeh et al. [26] combined SVM with the partial least square (PLS) to forecast the CO concentration in the region of Teheran. The proposed gray-box model achieves more accurate prediction as well as higher computational efficiency than single machine learning technique (black-box). We expect our study to provide a new possibility for modeling and simulation.

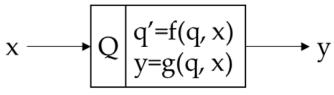
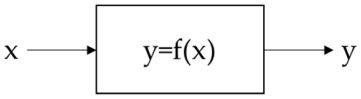
The rest of this paper is organized as follows. Section 2 presents the principle of the gray-box based modeling approach and two gray-box modeling methods of the ACD process. Two experiment cases combining the data and mechanism are elaborated in Sections 3 and 4, respectively. The results show that the proposed two approaches outperform conventional ones in model accuracy. Finally, the conclusions and future work are illustrated in Section 5.

2. Materials

2.1. The Gray-Box Based Modeling Approach

The mechanism-modeling and data-modeling both have their own advantages and limitations. Table 1 illustrates the generalized characteristics of two symmetric modeling approaches from different perspectives. From the point of view of the model representation and structure, the mechanism-modeling approach is to make a dynamic map of the inputs and states to the output variables based on cause–effect, and thereby this model requires knowledge of the target system. Conversely, the data-modeling approach merely associates the input variables with the output variables in static map form. As a result, researchers can build the data model without any prior knowledge. The former approach has been dealt with in research focusing on clear causality, such as physical and operational laws, while the latter has been studied in research areas of intelligent techniques, such as ANN and SVR.

Table 1. Comparison of the mechanism (white-box) model and data (black-box) model.

	Mechanism (White-Box) Model	**Data (Black-Box) Model**
Model representation	Cause-effect relationship between variables	Associational relationship between variables
Structure of the model	System knowledge required Dynamic map of (input, state) to output	No knowledge about system required Static map of input to output
	 (State Q within model)	 (No state within model)
Modeling means	Physical and/or operational laws	Intelligent techniques
Condition for valid prediction	Model validation	System structure remains unchanged before and after training
Anomaly/non-existing system	Applicable (as in rare event or new design)	Not applicable

The mechanism-modeling approach, which is based on knowledge, enables not only the prediction of the future under a different condition when training, but also the controlling and planning of the system. On the other hand, the data-modeling approach, which is built on learned information, enables to describe the past and to predict the future under the same condition when training. Such a difference allows the mechanism-modeling approach to conduct an analysis of a system with an abnormal or non-existing system, for example, a rare event of new design, which is impossible in

the data-modeling approach. However, as a real system has high complexity and various factors influencing it, the mechanism of a real system is hard to be totally obtained.

From the limitation and comparison of the mechanism modeling (white-box) and data modeling (black-box) approaches discussed above, we confirmed that it is difficult to describe a complex system using a single approach. As a consequence, it is necessary to propose a new modeling methodology that can combine the advantages of each and enhance the performance. It is also important to identify how the mechanism (white-box) model and data (black-box) model are constructed from prior knowledge and observation data, respectively, and then how these two models complement each other.

Focusing on the topic of system modeling, this paper firstly compares the conventional mechanism-modeling and data-modeling approaches, and then proposes the gray-box modeling approach which overcomes their limitations by combining the two modeling approaches. The modeling framework is depicted in Figure 1. The white-box modeling using the system mechanism can only model the system in the ideal experimental condition (the system mechanism is totally available). With regard to the black-box modeling approach, it is based on the data that can only describe the correlation of the system components and ignores the causality. Therefore, to deal with this problem, a modeling approach of the gray-box system should be established integrating the mechanism, dynamic and multi-granularity big data. In this paper, atmospheric dispersion is studied as the case of the typical gray-box system. Then, the two modeling approaches for this gray-box system are proposed.

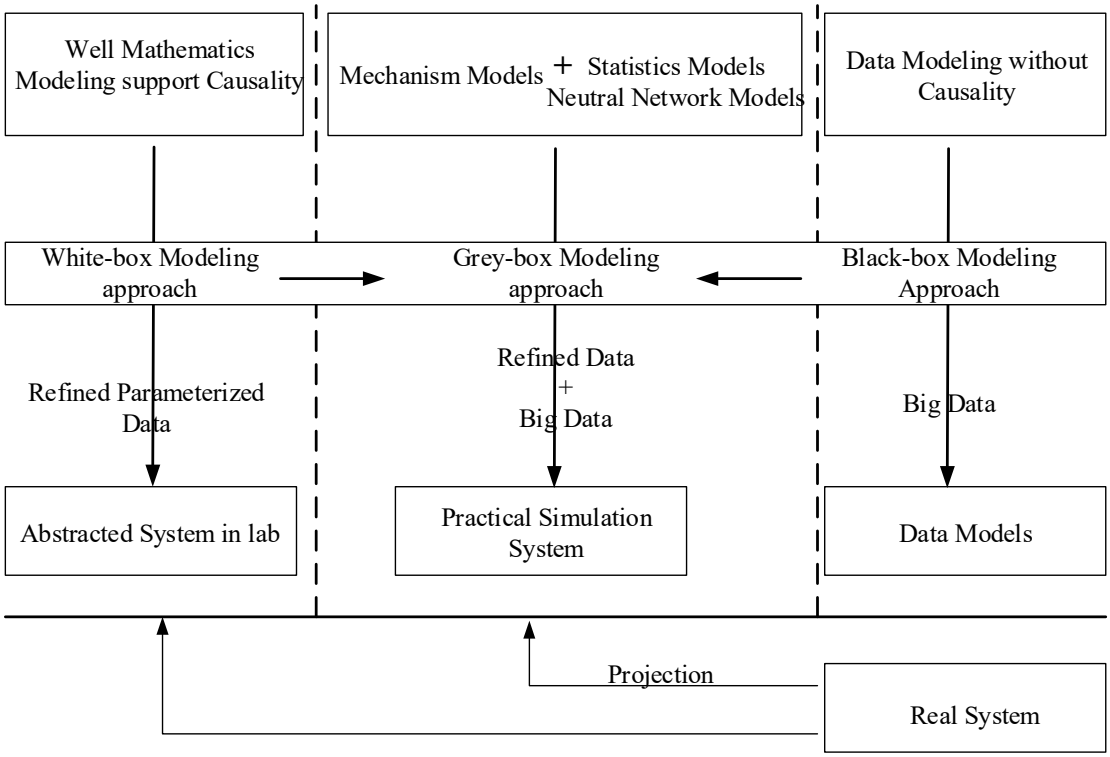


Figure 1. The Gray-box based Modeling Approach integrating both the Mechanism model and Data model.

2.2. Dynamic Data Driven Atmospheric Dispersion Modeling Method (from White-Box to Gray-Box)

The prediction performance of traditional atmospheric dispersion mechanism model depends on the correct setting of model parameters. However, the ACD is affected by many factors. These parameters usually changes dynamically in a field case. Therefore, it is difficult to make real-time and accurate observations of these parameters, especially in dynamic meteorological environments. Therefore, the simple mechanism-based modeling technique fails to give the prediction accurately

enough. As a result, it is necessary to model the ACD process in the manner of a gray-box modeling approach.

To address this problem, the data assimilation method is applied to introduce the dynamic observation into the Gaussian multi-puffs model. The introduction of the dynamic observation helps to correct and update the model parameters, and improve the prediction accuracy of the gray-box system consequently. The idea of gray-box modeling using data assimilation is shown in Figure 2.

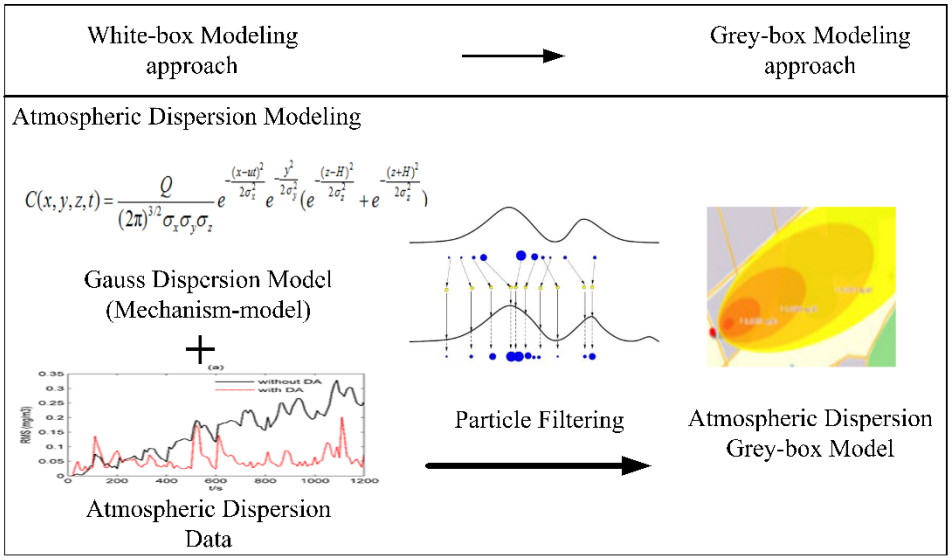


Figure 2. From White-box based Modeling to Gray-box Modeling: The Case of Atmospheric Dispersion Modeling.

2.2.1. Gaussian Dispersion Model

The Gaussian dispersion model is widely used in the modeling of atmospheric dispersion. In practice, some ACD models are also on the basis of the Gaussian model. Because the mechanism is relatively simple, this model predicts ACD process quickly. Taking the instant release of a point source as the case in this paper, the Gaussian puff model can be induced in Equation (1) according to the statistical theory:

$$C(x,y,z,t) = \frac{Q}{(2\pi)^{3/2}\sigma_x\sigma_y\sigma_z} e^{-\frac{(x-ut)^2}{2\sigma_x^2}} e^{-\frac{y^2}{2\sigma_y^2}} (e^{-\frac{(z-H)^2}{2\sigma_z^2}} + e^{-\frac{(z+H)^2}{2\sigma_z^2}}), \quad (1)$$

where x, y, and z are the coordinates in the downwind direction, the crosswind direction, and the vertical direction, respectively. t is the start time of emission, and H represents the height of source. Q is the total mass of atmospheric pollutants contained in puffs, and u is the average wind speed. σ_x, σ_y, and σ_z are the dispersion coefficients at different distance in x, y, and z directions, respectively. These dispersion coefficients are usually determined by the atmospheric stability level, which is commonly classified by the Pasquill–Gifford–Turner method [28,29]. A set of dispersion coefficient empirical formulas is [30]:

$$\begin{cases} \sigma_x = \sigma_y = ax^b \\ \sigma_z = cx^d \end{cases} \quad (2)$$

where a, b, c, d are influence factors. It can be seen that, in the Gaussian puff model, the puff propagates with the wind direction. The concentration of atmospheric contaminants obeys Gaussian distribution in x, y and z directions. When the source releases continuously, the continuously released plume can be regarded as a superposition of several puffs which are sequentially released at small time intervals.

Therefore, the concentration of atmospheric pollutants at an interest point in space is equal to the superposition of all released puffs:

$$C(t) = \sum_{i=1}^{n} f(t, z, l, q_j, t_{start} + (i-1)\delta) \tag{3}$$

The function $f(t, z, l, q, t_{start})$ represents the dispersion process of single puff. It calculates the concentration at an observation point of the position vector z at time t from the source whose releasing start-time is t_{start}, position vector is l and instantaneous release amount is q. q_j is the mass of pollutants released for the jth puff, which is equivalent to the release rate of source. n is the number of puffs, and δ is the interval time between releases of each puff. The concentration of atmospheric contaminants in dynamic source release rate and wind field scenarios can be approximated by function f.

2.2.2. Data Assimilation Framework Based on a Particle Filter

Based on the Gaussian multi-puffs model described above, the ACD process released from the continuous point source can be described by accurate parameter setting. However, the Gaussian model often fails to obtain accurate parameter values of a model timely in practical applications. For one thing, the monitoring devices can only give the average value of a specific parameter of ACD during a period of time. For example, only the average values of wind speed and wind direction can be obtained from meteorological stations. For another thing, the traditional Gaussian model is static, and its parameters always remain unchanged during prediction. Therefore, when the local meteorological conditions change dynamically, it is difficult to model the real ACD process only by a Gaussian multi-puffs model. The mechanism-based model cannot adapt to the dynamic meteorological environment. In order to solve this problem, the particle filter is used as a data assimilation method to construct a dynamic data-driven Gaussian dispersion model [31]. In this way, the mechanism model is corrected by the dynamic monitoring data. The system state (model parameters) is estimated by real-time monitoring data, so the model can be applied to the modeling of ACD process, which is a typical gray-box system.

The particle filter, also known as the Sequential Monte Carlo (SMC) method, is based on Bayesian inference and random sampling technique to recursively estimate the state of dynamic systems based on observed data [32,33]. The core idea is to use a series of weighted random sampling particles to approximate the posterior probability density function of the system state. Since particle filtering can estimate arbitrary probability density and has fewer assumptions about the model, it becomes an effective method for data assimilation of complex systems. The basic particle filtering algorithm consists of four repeated steps: initialization, importance sampling, weight update, and resampling. In order to apply particle filtering into data assimilation based on a Gaussian multi-puffs model, state space modeling on the dispersion of atmospheric contaminants must be performed. In general, a dynamic system can be described by a discrete state space model [34], including a state transition model (Equation (4)) and an observation model (Equation (5)):

$$s_{t+1} = f(s_t, t) + \gamma(t)\mathrm{d} \tag{4}$$

$$m_t = g(s_t, t) + \omega(t) \tag{5}$$

where s_t and m_t represent the system state variable observed time t, respectively. The function f in the state transition model describes the evolution of the system state over time, while the function g in the observation model defines the relationship between system state and observation values. γ and ω are independent random variables, describing system state noise and observed noise, respectively. Here, the influence coefficients a, b, c, d in Gaussian multi-puffs model expression are selected as the system state, which are determined by atmospheric stability (affected by meteorological conditions such as solar radiation intensity, wind field, cloud cover, etc.). Due to the influence of various meteorological factors, it is difficult to make accurate predictions in real time. At the same time, the influence coefficients

determine that the concentration of atmospheric contaminants follows Gaussian distribution, so they are selected as the system state. In terms of the state transition equation, the change of the level of atmospheric stability is usually slow in the actual environment. In addition, the law of change is not clear, which is difficult to be modeled by mathematical equations. Therefore, we select an identity function to denote the state transfer function f, and use the state noise γ (set as Gaussian state noise in this paper) to realize the transition and evolution of the system state in each time step. In the observation model, since the ACD process is a dynamic system, we can build the relationship between system state and observation (value of ACD) though Gaussian multi-puffs model (as a function g). The observed noise set as Gaussian white noise in the observation model is usually derived from the observation device itself in reality. By applying dynamic observation data, a gray-box model combining data and mechanism through the particle filtering technique is established. It is suitable for modeling an atmospheric dispersion process (a gray-box system mentioned before) in a dynamic environment.

2.3. Atmospheric Dispersion Modeling Method Based on Gaussian-Machine Learning (from Black-Box to Gray-Box)

Other than the mechanism-modeling, data-modeling approaches like machine learning for black-box system is also widely used in atmospheric dispersion [24,35,36]. Unlike the mechanism model, the machine learning method builds a mapping relationship between specific parameters and dispersion concentration through a predefined training set [37]. The training sets are always based on some datasets of real experiments, e.g., the Prairie Grass dataset [38] and the Indianapolis dataset [39]. Furthermore, the dispersion of atmospheric contaminants has been studied by scientists and lots of dispersion mechanisms are proposed. Machine learning models without the consideration of mechanism and prior knowledge often give poor performances due to the complex mechanisms of the real atmospheric dispersion. Therefore, constructing the gray-box model by adding some mechanisms into the data model is a feasible way to solve this problem.

2.3.1. Support Vector Regression

Common used machine learning methods for atmospheric dispersion modeling and prediction include ANN, SVR, and so on [27]. Here, SVR is used to construct the atmospheric dispersion model. Different from the SVM model used in classification problems [40], SVR always deals with the regression problems. SVR fits a complex function relationship by mapping input data to high-dimensional feature space and performs the linear regression. The remarkable characteristic of SVR is that its training goal is not to minimize the prediction error, but to make the model more generalized by minimizing the generalization error bound [40]. Given a training set $\{(x_1, z_1), \ldots, (x_l, z_l)\}$, where $x_i \in R^n$ is input and $z_i \in R^l$ is output, the standard form of SVR can be expressed as:

$$\begin{gathered} \min_{\omega,b,\xi,\xi^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \\ s.t.\ \omega^T\phi(x_i) + b - z_i \le \varepsilon + \xi_i, \\ z_i - \omega^T\phi(x_i) - b \le \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, l, \end{gathered} \tag{6}$$

where C is the regularization parameter, ε is the error tolerance, ξ and ξ * represent slack variables respectively. The fitting function is listed below:

$$y(x) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x) + b \tag{7}$$

where $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ represents kernel function, the commonly used Radial Basis Function (RBF) is chosen as the kernel function of SVR in this paper. α is the support vector. In atmospheric

dispersion modeling problems, $x_i \in R^n$ represents the related parameter of atmospheric dispersion, and the output $y(x)$ of SVR is the concentration of the interest points. A series of parameters related to atmospheric dispersion are usually selected as input features, as shown in Table 2. The regularization parameter C and the expansion coefficient σ in RBF function affect the complexity and the performance of the model. In order to construct an optimal SVR model, the optimal combination of parameter values will be selected according to the model performance during model construction process.

Table 2. Common parameters in atmospheric dispersion.

Parameters	Symbol	Unit	Whether Choosing as an Input Parameter
Downwind distance	D_x	m	Y
Crosswind distance	D_y	m	Y
Source height	H	m	N
Interest point height	z	m	Y
Source release rate	q	g s^{-1}	Y
Atmospheric stability level	STA	/	Y
Wind direction	d	deg	Y
Wind speed	v	m s^{-1}	Y
Mixed layer height	z_m	m	N
Cloud height	z_c	m	N
Cloud cover rate	P_c	%	N
Temperature	T	K	N

2.3.2. Feature Construction Method Based on Gaussian Model Knowledge

In the construction of SVR model, input parameters are some original observation parameters in atmospheric dispersion scenes. These parameters are easy to obtain in actual scenes, and their quantity and quality are also guaranteed. However, due to the complexity of ACD process, the relationship between many observation parameters and the concentration of interest points is complex and difficult to describe. The complex mapping relationship between input and output brings some difficulties to the training of SVR model, which in turn affects the accuracy of the model. To solve this problem, more efficient input features should be proposed to reduce the training difficulty. The ACD is affected by many factors, so it is difficult to get a high-accurate model. However, ACD is not a complete black-box system. Therefore, the knowledge of mechanism models can be introduced into the feature construction of data model such as SVR. As mentioned in Section 2, Gaussian dispersion model, which is a classical atmospheric dispersion model, can simulate the dispersion of atmospheric pollutants in many scenarios effectively. Moreover, the form of Gaussian dispersion model is simple and fast to calculate. Therefore, the knowledge of Gaussian multi-puffs model is applied to the feature construction of the SVR model in this paper. Two items of Gaussian multi-puffs model G_y, G_z are selected and then added into the input features of SVR model. The expression is as follows:

$$\begin{cases} G_y = exp(-\frac{D_y^2}{2\sigma_y^2}) \\ G_z = exp[-\frac{(z+H)^2}{2\sigma_z^2}] + exp[-\frac{(z-H)^2}{2\sigma_z^2}] \end{cases} \tag{8}$$

where G_y, G_z represent the dispersion coefficient at different distances in the y, z directions. Gaussian parameters described above combine many factors such as wind speed, wind direction, downwind distance, crosswind distance and atmospheric stability level. It is a direct and efficient way to describe the dispersion of atmospheric contaminants. Compared with the original observation parameters, Gaussian parameters G_y, G_z are high-dimensional features, which can reduce the complexity of the input–output mapping relationship effectively. Thus, they are used to construct the Gaussian-SVR model which is a gray-box model for the prediction of ACD. The idea is shown in Figure 3.

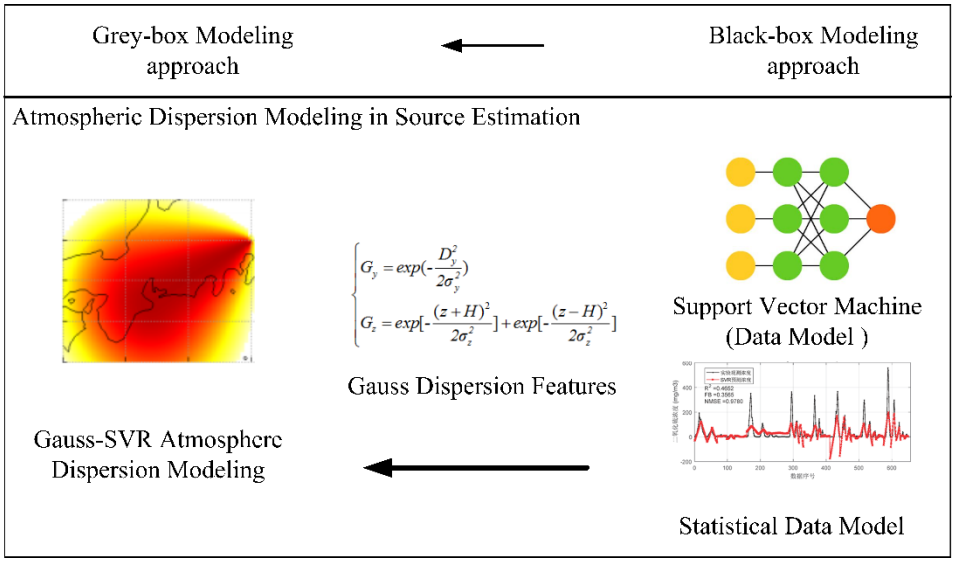


Figure 3. From Black-box based Modeling to Gray-box Modeling: The case of Atmospheric Dispersion Modeling in Source Estimation.

3. Case 1: Dynamic Data Driven Atmospheric Dispersion Modeling Method (from White-Box to Gray-Box)

3.1. Experimental Design

We constructed a simulated atmospheric pollutant emission and dispersion scenario in a commercial process hazard analysis software (PHAST) [41] to verify the prediction performance of abovementioned dynamic data-driven Gaussian multi-puffs model in dynamic meteorological conditions. In this emission scenario, the study area is a square area of 1000 × 1000 m^2. The source is located at (0, 0, 50), from which puffs are released at an interval of 10 s throughout the simulation. The release rate of the source, which is also the mass of atmospheric pollutants contained in puffs, is set to a random variable with a mean of 50 g and a standard deviation of 5 g (10% mean). The wind field parameters are modeled as Gaussian white noise. The wind speed obeys a Gaussian distribution with a mean of 3 m/s and a standard deviation of 0.3 m/s, while the wind direction obeys a Gaussian distribution whose mean is 220 degrees and standard deviation is 10 degrees. In order to construct a dynamic meteorological condition, the atmospheric stability level [42,43] is set as changing dynamically with time (shown in Table 3). The dynamic atmospheric stability level will affect the influence coefficients in the Gaussian model. The influence coefficients change linearly in the three time periods of 0–400 s, 400–800 s, and 800–1200 s. Using this simulation scenario, the dispersion of atmospheric contaminants at a height of 30 m is simulated based on a Gaussian multi-puffs model. The simulation time is set as 1200 s.

Table 3. Values of atmospheric stability and influence coefficients in simulated dispersion scenarios.

Time (s)	Atmospheric Stability	Influence Coefficient			
		a	*b*	*c*	*d*
0	Level A	0.23	1.00	0.10	1.16
400	Level B	0.23	0.97	0.16	1.02
800	Level C	0.22	0.94	0.25	0.89
1200	Level D	0.22	0.91	0.40	0.76

As shown in Table 4, the control experiments are used to illustrate the modeling effect of the data assimilation model in dynamic meteorological conditions. The control group uses traditional

Gaussian multi-puffs model as a dispersion model, representing the white-box modeling approach using mechanisms. Moreover, fixed source terms, wind field parameters, and atmospheric stability level are used as estimates for the corresponding parameters in the actual environment. The mass of atmospheric pollutants contained in all puffs is set to 50 g. The wind direction is 220 degrees and the wind speed is 3 m/s. The atmospheric stability level influence coefficients are also set in Table 3. In contrast, the dynamic data-driven Gaussian multi-puffs model is used in the experimental group. Driven by dynamic monitoring data, the dynamic correction and estimation of the influence coefficients of the model are realized. The dynamic monitoring data is acquired by several virtual unmanned aerial vehicles (UAVs) along particular paths in simulated dispersion scenarios which is detailed in reference [44]. The source release rate and wind field settings in the experimental group are the same as in the control group.

Table 4. Experimental design.

Experiment Number	Atmospheric Dispersion Model	Description
A	Gaussian multi-puffs model (white-box model)	Control group: comparison with experiment B
B	Gaussian multi-puffs model with data assimilation (gray-box model)	Experimental group: test modeling effect of data assimilation

The experiments are carried out as follows: Firstly, the emission and dispersion of atmospheric contaminants under dynamic meteorological conditions are simulated. The monitoring concentration data is collected by the virtual UAVs. In the experimental group, with the input of the monitoring data, the system state (influence coefficient) is corrected and estimated dynamically in the framework of data assimilation. As a result, the prediction of the concentration is obtained. In the control group, the dispersion of atmospheric contaminants is predicted directly according to the traditional Gaussian multi-puffs model.

3.2. Experimental Results

The variations of four influence coefficients during the experiments are shown in Figure 4. It can be seen that, in simulated dispersion scenarios, four coefficients are changing with time dynamically. In the control group, since the traditional Gaussian multi-puffs model is static during running process, its model parameters always stay at the initial values. The values of influence coefficients in the model deviate from the initial setting values over time gradually. In contrast, the influence coefficients in the experimental group are corrected and updated dynamically due to the introduction of dynamic monitoring data. With the help of data assimilation, the influence coefficients in the experimental group model are close to the initial setting values. Moreover, the change of influence coefficients also follow similar trends, as shown in Figure 4. In order to compare the performance of proposed model in concentration prediction, the errors at the observation point of the two groups of experiments are calculated. The results are shown in Figure 5. The figure shows that the root-mean-square error (RMSE) of dynamic data driven Gaussian multi-puffs model is significantly lower than that of the traditional Gaussian multi-puffs model in concentration prediction. The error of the control group is due to the fixed influence coefficients in the model. Because the influence coefficients change with time, the error of model accumulates over time, resulting in the rise of RMSE. In comparison, the experimental group guarantees the accuracy of the model by correcting the state of the system with the help of dynamic monitoring data.

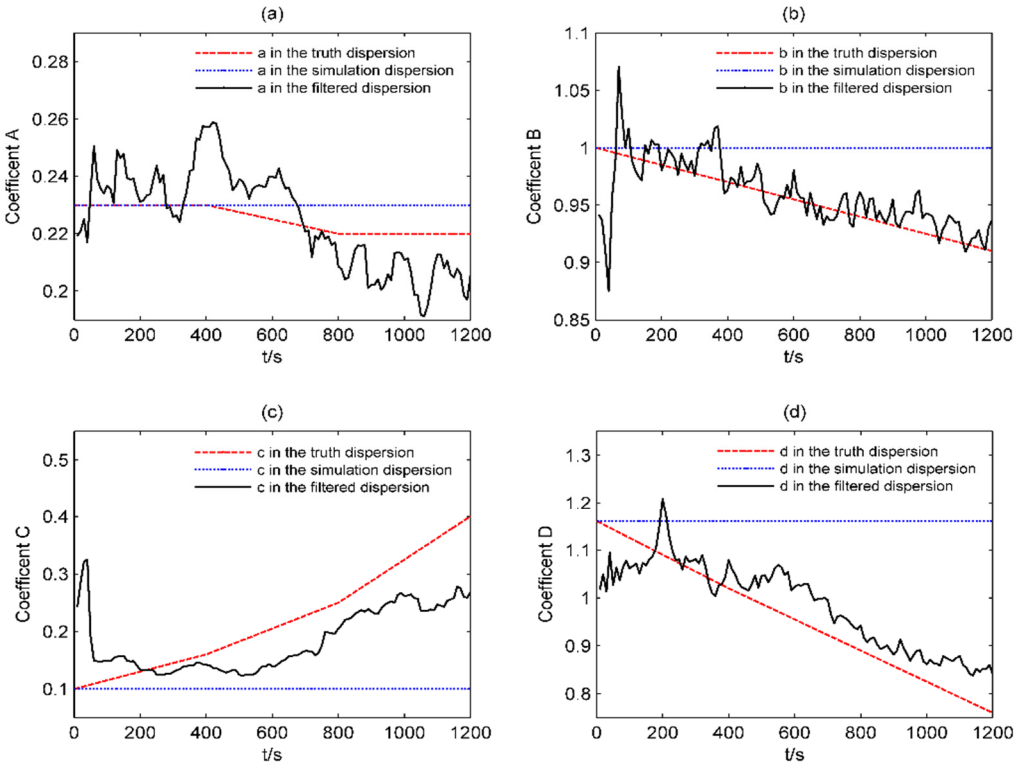


Figure 4. The comparisons of dispersion parameters in experiments. (**a**) Values of coefficient A; (**b**) Values of coefficient B; (**c**) Values of coefficient C; (**d**) Values of coefficient D.

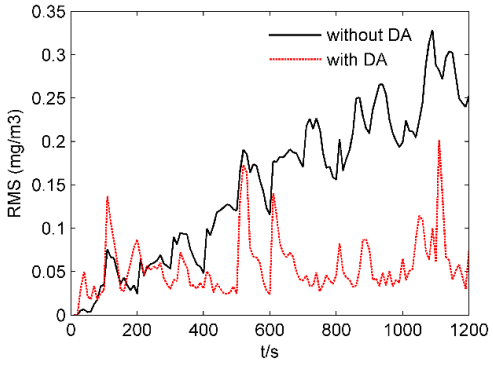


Figure 5. The results of concentration prediction.

The effectiveness of the dynamic data-driven Gaussian multi-puffs model can be validated by the comparison of the error distribution. The error distribution shows the prediction of concentration over the area of ACD. As shown in Figure 6, the concentration prediction based on the dynamic data-driven Gaussian multi-puffs model is significantly better than traditional Gaussian multi-puffs model in most areas. According to the discussion in Section 1, with the support of monitoring data, the dynamic data-driven Gaussian multi-puffs model is built by a gray-box modeling approach. The case testifies the better performance and effectiveness of gray-box model.

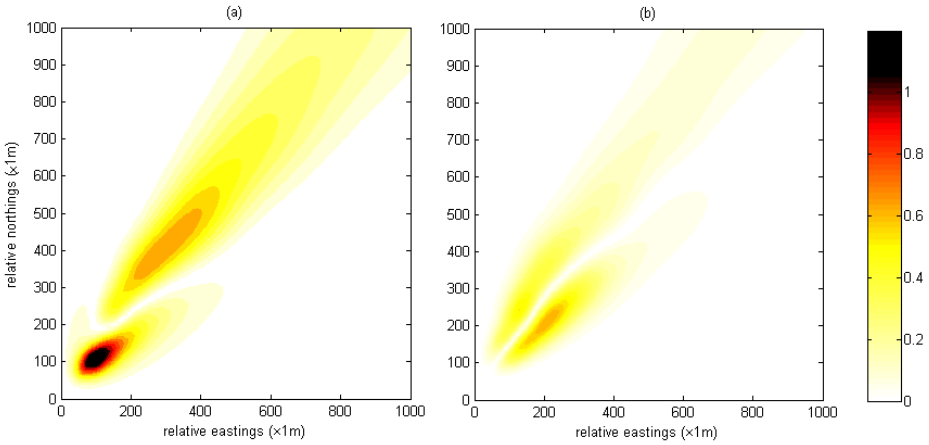
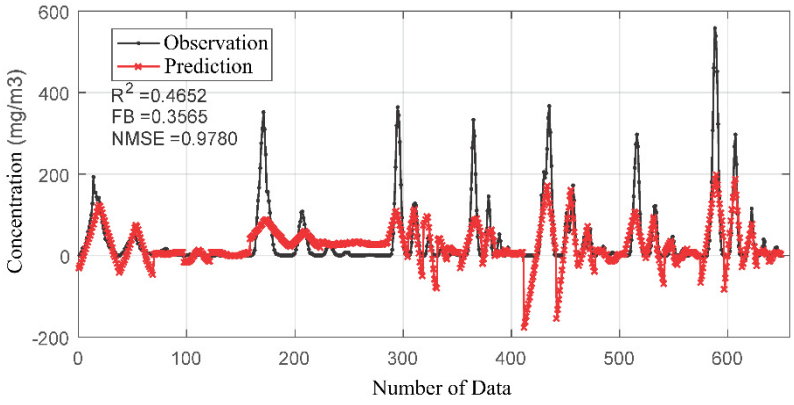


Figure 6. The error distribution of t = 1200 s. (**a**) Error distribution of traditional Gaussian multi-puffs model; (**b**) Error distribution of dynamic data-driven Gaussian multi-puffs model.

4. Case 2: Atmospheric Dispersion Modeling Method Based on Gaussian-Machine Learning (from Black-Box to Gray-Box)

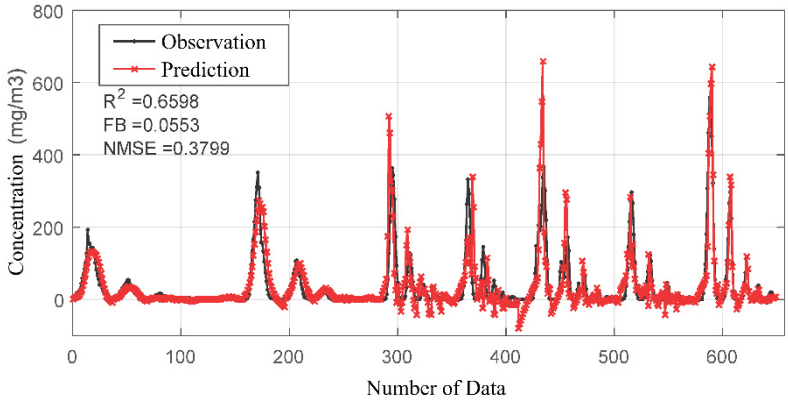
In order to verify the effectiveness of the proposed Gaussian-SVR model in the prediction of ACD, this paper compares the SVR model which is based on the original observation parameters with the Gaussian-SVR model in the Prairie Grass dataset. The Prairie Grass dataset is widely used in the field of atmospheric dispersion [27,45,46]. This tracer experiment was carried out in an open country in O'Neil, NE (USA, 42.49°N and 98.57°W) from July to August, 1956. The sulfur dioxide (SO_2) tracer was released from a continuous point source at the height of 0.46 m without buoyancy. Concentration data were collected by five semi-circular arcs of receptors. There are 68 releases containing tracer data (6888 valid samples used in this paper) and meteorological data in the data set.

The first 60 release experiments data (a total of 6239 observation concentration samples) in the Prairie Grass dataset are used as training and verification set, and the remaining eight pieces of release data (a total of 649 observation concentration samples) are used as a test set. The construction and training of the SVR model is conducted by LIBSVM [47]. The optimal value of two hyper parameters in the SVR model is determined by the cross-validation method. Specifically, Gaussian parameters are integrated with the SVR (black-box) model as mechanism knowledge to build the Gaussian-SVR (gray-box) model. Subsequently, the original-SVR model and Gaussian-SVR model are tested on the test set, respectively. The results are shown in Figures 7 and 8.



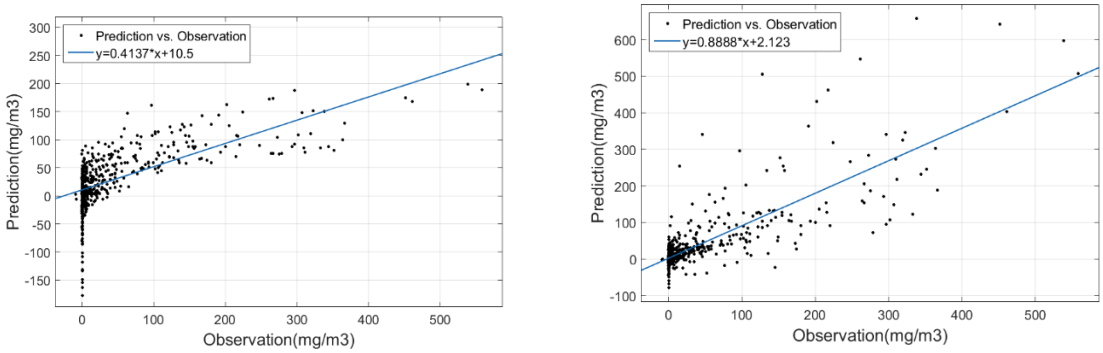
(**a**) original-SVR model

Figure 7. *Cont.*



(**b**) Gaussian-SVR model

Figure 7. The results of SVR based Atmosphere Dispersion model: Gauss Model-SVR model.



(**a**) original-SVR model (**b**) Gaussian-SVR model

Figure 8. Observation data and Prediction data fitting curve: Gauss Model-SVR model.

Figure 7a shows that the prediction accuracy of the original-SVR model is not satisfying, especially in high concentration values (above 200 mg/m^3) prediction, which is far lower than the experiment observations. In the meantime, some negative values appear in the predictions of original-SVR model, which is inconsistent with the actual situation obviously. In contrast, Figure 7b indicates that the model predictions of Gaussian-SVR model are closer to the experiment observations and have better accuracy in the prediction of high concentration values. In addition, Figures 7 and 8 both show that the negative values in Gaussian-SVR model predictions are reduced obviously. Some model evaluation indexes are used to measure the performance of prediction models, such as the correlation coefficient squared (R^2), the score deviation FB, and the normalized mean square error (NMSE) [24,27,41]. These three indexes are calculated and shown in Figure 7. Obviously, the prediction coefficient R^2 of Gaussian-SVR (0.6598) is significantly higher than that of original-SVR (0.4652). Furthermore, the score deviation FB and the normalized mean square error NMSE of Gaussian-SVR predictions (0.0553 and 0.3799) are also lower than original-SVR (0.3565 and 0.9780). Moreover, the fitting curve is also applied by many researchers to evaluate the accuracy of prediction data overall [45,48], which is exhibited in Figure 8. The linear fitting curve of Gaussian-SVR model is more close to "y = x" than that of original-SVR model clearly, indicating the prediction data of Gaussian-SVR (gray-box) model is more accurate.

Through the comparison of the results above, it can be concluded that the prediction performance of the SVR model is significantly improved by introducing the mechanism model knowledge into the feature construction of the SVR model. This shows that Gaussian features constructed by the mechanism model knowledge are more efficient than original parameters, which can reduce the training difficulty of the model effectively and improve the accuracy of the model.

5. Conclusions and Expectations

With the requirements of system analysis and prediction for human society, modeling and simulation techniques are widely used in many fields gradually, including scientific research, engineering practice, political economy, and so on. As a result, researchers are confronted with high demands on accurate system modeling. Existing mechanism-modeling and data-modeling approaches have good performance in describing white-box and black-box systems, respectively. However, most systems faced in the practical scenarios are gray-box systems. On this occasion, there are inevitable drawbacks and limitations in white-box modeling and black-box modeling. After the comparison of these two symmetric modeling approaches, we integrate these two approaches to obtain a gray-box modeling method which can well describe the gray-box systems. The gray-box modeling approach not only considers the prior knowledge of the target system, but also applies the intelligent statistical as well as machine learning techniques. Taking the typical gray-box system cases of ACD as examples, this paper demonstrates two symmetric gray-box modeling methods which combine both mechanism and data. For the problem in which the mechanism (white-box) model is difficult to model the atmospheric dispersion model in a dynamic meteorological environment, the dynamic monitoring data are served as inputs into the Gaussian multi-puffs model. Through the data assimilation method, the dynamic correction of the model parameters is realized. As a result, the static Gaussian multi-puffs model can be adapted to the modeling of the dynamic dispersion of atmospheric contaminants. For the problem in which the data (black-box) model of atmospheric dispersion prediction is not accurate enough, the knowledge of a mechanism model is introduced into the feature construction of an SVR model. The mechanism knowledge about Gaussian model is used to reduce the difficulty of model training, thus improving the prediction accuracy of the model.

As a main contribution, we provide a possibility of a new symmetric modeling approach which is superior to a single mechanism (white-box) modeling approach or a data (black-box) modeling approach. However, a gray-box modeling approach is only a methodology. When facing different real systems, various prior knowledge and data models can be combined together. Moreover, even the same target system can integrate different prior knowledge and data models. On this occasion, finding effective mechanism models and data models is a great challenge.

In the future, further research on the modeling of ACD in emergency management will be carried out to promote the integration of data and mechanism modeling approaches. This integration is reflected in many aspects. Firstly, the integration of multi-sources data, which means obtaining more accurate meteorological and monitoring data according to the observation of various monitoring resources—secondly, further integration of data models and mechanism models. For example, aiming at the problem of the poor accuracy for the Gaussian-SVR model in complex terrain, the mechanism models (such as CFD) are further combined to design input features for higher precision modeling.

Author Contributions: According to symmetric modeling and simulation, a gray-box based modeling approach integrating both mechanism-model and data-model is proposed in this paper. Generally, mechanism-models (white-box) and data-models (black-box) are used to model different target systems. However, the former model is suitable for the structural and diagnostic analysis of system, but unpractical to describe a complex real system totally. While the latter model can build a model thoroughly bypassing the system mechanism. It describes the correlation of the data, but cannot explain the relationship between the input and the output interpretable. Moreover, the data-modeling approach cannot cope with anomalies and changing circumstances of the system. To mitigate the disadvantages of single white-box or black-box approach, a new gray-box modeling methodology integrating both mechanism-model and data-model is proposed. As a typical gray-box system, the air contaminant dispersion (ACD) is taken as an example. Two symmetric experiment cases combining the data and mechanism are elaborated in this paper. The results show that the gray-box based modeling approach achieves more accurate prediction as well as higher computational efficiency than single modeling approach (white-box or black-box). B.C. and Y.W. completed the method and concept parts of this article. B.C. and R.W. review the experimental analysis of this article. Z.Z., L.M., X.Q. and W.D. jointly completed the writing of the article. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research & Development (R&D) Plan under Grant No. 2018YFC0806900 and the National Natural Science Foundation of China under Grant Nos. 71673292, 21808181,

61673388, 91646101, and the National Social Science Foundation of China under Grant No. 17CGL047 and Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion.

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

References

1. Chen, B.; Qiu, X.; Wang, Y. An Intelligent ACP based Experimental Approach. *J. Syst. Simul.* **2017**, *29*, 2064–2072.
2. Kedi, H. *System Simulation Techniques*; Press of National University of Defense Technology: Changsha, China, 1998.
3. Gerstlauer, A.; Haubelt, C.; Pimentel, A.D.; Stefanov, T.P.; Gajski, D.D.; Teich, J. Electronic system-level synthesis methodologies. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2009**, *28*, 1517–1530. [CrossRef]
4. Builder, C.H.; Bankes, S.C. *Artificial Societies: A Concept for Basic Research on the Societal Impacts of Information Technology*; RAND Corporation: Santa Monica, CA, USA, 1991.
5. Yi, L.; Shunjiang, N.; Wenguo, W. Development of the Public Safety System and a Security-Guaranteed Society. *Strateg. Study Chin. Acad. Eng.* **2017**, *19*, 118–123.
6. Bock, H.G.; Carraro, T.; Jäger, W.; Körkel, S.; Rannacher, R.; Schlöder, J.P. *Model Based Parameter Estimation: Theory and Applications*; Springer Science & Business Media: Berlin, Germany, 2013; Volume 4.
7. Zhu, Z.; Chen, B.; Qiu, S.; Wang, R.; Wang, Y.; Ma, L.; Qiu, X. A data-driven approach for optimal design of integrated air quality monitoring network in a chemical cluster. *R. Soc. Open Sci.* **2018**, *5*, 180889. [CrossRef] [PubMed]
8. Chen, Y.L.; Chen, J.M.; Tung, C.W. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. *Decis. Support Syst.* **2006**, *42*, 1503–1520. [CrossRef]
9. Sagae, K.; Lavie, A. Combining rule-based and data-driven techniques for grammatical relation extraction in spoken language. In Proceedings of the Eighth International Conference on Parsing Technologies, Nancy, France, 23–25 April 2003.
10. Dahl, N.; Xue, H.; Hu, X.; Xue, M. Coupled fire–atmosphere modeling of wildland fire spread using DEVS-FIRE and ARPS. *Nat. Hazards* **2015**, *77*, 1013–1035. [CrossRef]
11. Wilkie, D.; Sewall, J.; Lin, M.C. Transforming GIS data into functional road models for large-scale traffic simulation. *IEEE Trans. Vis. Comput. Graph.* **2011**, *18*, 890–901. [CrossRef]
12. Varma, D.R.; Guest, I. The Bhopal accident and methyl isocyanate toxicity. *J. Toxicol. Environ. Health* **1993**, *40*, 513–529. [CrossRef]
13. Fernando, H.; Lee, S.; Anderson, J.; Princevac, M.; Pardyjak, E.; Grossman Clarke, S. Urban fluid mechanics: Air circulation and contaminant dispersion in cities. *Environ. Fluid Mech.* **2001**, *1*, 107–164. [CrossRef]
14. Turner, D.B. A diffusion model for an urban area. *J. Appl. Meteorol.* **1964**, *3*, 83–91. [CrossRef]
15. Pontiggia, M.; Derudi, M.; Busini, V.; Rota, R. Hazardous gas dispersion: A CFD model accounting for atmospheric stability classes. *J. Hazard. Mater.* **2009**, *171*, 739–747. [CrossRef]
16. Xing, J.; Liu, Z.; Huang, P.; Feng, C.; Zhou, Y.; Zhang, D.; Wang, F. Experimental and numerical study of the dispersion of carbon dioxide plume. *J. Hazard. Mater.* **2013**, *256–257*, 40–48. [CrossRef] [PubMed]
17. Flesch, T.K.; Wilson, J.D.; Yee, E. Backward-time lagrangian stochastic dispersion models and their application to estimate gaseous emissions. *J. Appl. Meteorol.* **1995**, *34*, 1320–1332. [CrossRef]
18. Wilson, J.D.; Sawford, B.L. Review of Lagrangian stochastic models for trajectories in the turbulent atmosphere. *Bound.-Layer Meteorol.* **1996**, *78*, 191–210. [CrossRef]
19. Briggs, G. *Diffusion Estimation for Small Emissions. Preliminary Report*; Atmospheric Turbulence and Diffusion Lab., National Oceanic and Atmospheric Administration: Oak Ridge, TN, USA, 1973.
20. Hanna, S.R.; Briggs, G.A.; Hosker, R.P., Jr. *Handbook on Atmospheric Diffusion*; Atmospheric Turbulence and Diffusion Lab., National Oceanic and Atmospheric Administration: Oak Ridge, TN, USA, 1982.
21. Krysta, M.; Bocquet, M.; Sportisse, B.; Isnard, O. Data assimilation for short-range dispersion of radionuclides: An application to wind tunnel data. *Atmos. Environ.* **2006**, *40*, 7267–7279. [CrossRef]
22. Reddy, K.V.U.; Cheng, Y.; Singh, T.; Scott, P.D. Data assimilation in variable dimension dispersion models using particle filters. In Proceedings of the 2007 10th International Conference on Information Fusion, Quebec City, QC, Canada, 9–12 July 2007.

23. Zheng, D.; Leung, J.; Lee, B.; Lam, H. Data assimilation in the atmospheric dispersion model for nuclear accident assessments. *Atmos. Environ.* **2007**, *41*, 2438–2446. [CrossRef]
24. Pelliccioni, A.; Tirabassi, T. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. *Environ. Model. Softw.* **2006**, *21*, 539–546. [CrossRef]
25. Wang, B.; Chen, B.; Zhao, J. The real-time estimation of hazardous gas dispersion by the integration of gas detectors, neural network and gas dispersion models. *J. Hazard. Mater.* **2015**, *300*, 433–442. [CrossRef]
26. Yeganeh, B.; Motlagh, M.S.P.; Rashidi, Y.; Kamalan, H. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmos. Environ.* **2012**, *55*, 357–365. [CrossRef]
27. Ma, D.; Zhang, Z. Contaminant dispersion prediction and source estimation with integrated Gaussian-machine learning network model for point source emission in atmosphere. *J. Hazard. Mater.* **2016**, *311*, 237–245. [CrossRef]
28. Gifford, F.A., Jr. Use of routine meteorological observations for estimating atmospheric dispersion. *Nucl. Saf.* **1961**, *2*, 47–51.
29. Pasquill, F. The estimation of the dispersion of windborne material. *Met. Mag.* **1961**, *90*, 33.
30. Carrascal, M.; Puigcerver, M.; Puig, P. Sensitivity of Gaussian plume model to dispersion specifications. *Theor. Appl. Climatol.* **1993**, *48*, 147–157. [CrossRef]
31. Zhu, Z.; Qiu, S.; Chen, B.; Wang, R.; Qiu, X. Data-driven hazardous gas dispersion modeling using the integration of particle filtering and error propagation detection. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1640. [CrossRef] [PubMed]
32. Bouttier, F.; Courtier, P. Data assimilation concepts and methods March 1999. *Meteorol. Train. Course Lect. Ser. ECMWF* **2002**, *718*, 59.
33. Gordon, N.J.; Salmond, D.J.; Smith, A.F. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F-Radar Signal Process.* **1993**, *140*, 107–113. [CrossRef]
34. Senne, K. Stochastic processes and filtering theory. *IEEE Trans. Autom. Control* **1972**, *17*, 752–753. [CrossRef]
35. Boznar, M.; Lesjak, M.; Mlakar, P. A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. *Atmos. Environ. Part B Urban Atmos.* **1993**, *27*, 221–230. [CrossRef]
36. Krasnopolsky, V.M.; Schiller, H. Some neural network applications in environmental sciences. Part I: Forward and inverse problems in geophysical remote measurements. *Neural Netw.* **2003**, *16*, 321–334. [CrossRef]
37. Qiu, S.; Chen, B.; Wang, R.; Zhu, Z.; Wang, Y.; Qiu, X. Estimating contaminant source in chemical industry park using UAV-based monitoring platform, artificial neural network and atmospheric dispersion simulation. *RSC Adv.* **2017**, *7*, 39726–39738. [CrossRef]
38. Barad, M.L. *Project Prairie Grass, a Field Program in Diffusion*; Air Force Cambridge Research Center: Bedford, MA, USA, 1958; Volume 1.
39. Steven Hanna, J.; Olesen, H.R. *Indianapolis Tracer Data and Meteorological Data*; National Environmental Research Institute: Roskilde, Denmark, 2005.
40. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
41. Wang, R.; Chen, B.; Qiu, S.; Ma, L.; Zhu, Z.; Wang, Y.; Qiu, X. Hazardous source estimation using an artificial neural network, particle swarm optimization and a simulated annealing algorithm. *Atmosphere* **2018**, *9*, 119. [CrossRef]
42. Cervone, G.; Franzese, P. Non-Darwinian evolution for the source detection of atmospheric releases. *Atmos. Environ.* **2011**, *45*, 4497–4506. [CrossRef]
43. Wang, Y.; Huang, H.; Huang, L.; Ristic, B. Evaluation of Bayesian source estimation methods with Prairie Grass observations and Gaussian plume model: A comparison of likelihood functions and distance measures. *Atmos. Environ.* **2017**, *152*, 519–530. [CrossRef]
44. Wang, R.; Chen, B.; Qiu, S.; Zhu, Z.; Ma, L.; Qiu, X.; Duan, W. Real-Time data driven simulation of air contaminant dispersion using particle filter and UAV sensory system. In Proceedings of the 2017 IEEE/ACM 21st International Symposium on Distributed Simulation and Real Time Applications (DS-RT), Rome, Italy, 18–20 October 2017; pp. 1–4.
45. Cui, J.; Lang, J.; Chen, T.; Cheng, S.; Shen, Z.; Mao, S. Investigating the impacts of atmospheric diffusion conditions on source parameter identification based on an optimized inverse modelling method. *Atmos. Environ.* **2019**, *205*, 19–29. [CrossRef]

46. Ma, D.; Gao, J.; Zhang, Z.; Wang, Q. An Improved Firefly Algorithm for Gas Emission Source Parameter Estimation in Atmosphere. *IEEE Access* **2019**, *7*, 111923–111930. [CrossRef]
47. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]
48. Ma, D.; Deng, J.; Zhang, Z. Comparison and improvements of optimization methods for gas emission source identification. *Atmos. Environ.* **2013**, *81*, 188–198. [CrossRef]